Fitting logistic regression models

QERM 514 - Homework 4

22 May 2026

Background

This week's homework assignment focuses on fitting and evaluating logistic regression models. The data come from a unique tagging program for salmon in the Columbia River basin. Beginning in the mid-1990s, individual juvenile salmon have been captured in their natal rearing habitats in Idaho, Oregon, and Washington, and implanted with a passive integrated transponder (PIT) tag. These tagged fish can then be detected at numerous locations during their downstream migration to the sea, including most of the hydroelectric dams they pass, which allows researchers to estimate their survival. Those juveniles that survive for 1-4 years in the ocean before maturing can also be detected as they swim upstream towards their spawning grounds.

Many of these salmon belong to populations that are listed as threatened or endangered under the Endangered Species Act. As such, there is great interest in trying to understand how hydropower operations affect the survival of both juveniles and adults. In particular, the estimated smolt-to-adult returns (SARs) have been a focus due to the perceived delayed effects of the juveniles' downstream journey on their subsequent survival (so-called "delayed mortality"). Furthermore, previous work by Scheuerell et al. (2009) showed nonlinear relationships between SARs and migration timing within a year, indicating a possible window of opportunity for some fish and a mismatch with the environment for those migrating relatively early or late in the season.

Your assignment is to investigate how daily estimates of SARs for Chinook salmon from the Snake River basin vary across a portion of their migration season for one year, and explore whether there is any relationship between SARs and water temperature. The accompanying data file `srss_chin_sar.csv` contains information about tagged salmon detected at Bonneville dam (BON), the last dam juveniles pass as they head to sea and the first dam adults encounter upon their return. Here are descriptions of the fields of information.

- `day`: the day of the month in May (1-31)
- `smolts`: the number of tagged juveniles detected at BON on a given day
- `adults`: the number of surviving adults subsequently detected at BON
- `temp`: the water temperature recorded at BON on that day

Questions

a) Identify the three components of a GLM that you will need to fit a logistic regression model for survival given these data. (3 pts)

b) Plot daily estimates of survival against `day` and `temp` and describe any patterns you see. (3 pts)

c) Would it be reasonable to include both `day` and `temp` as predictors in the same model? Why or why not? (4 pts)

d) Fit a logistic regression model with survival as a function of only an intercept and compute the R^2 value. Based upon this model, what is the estimated mean survival for the month of May? Plot the model residuals against `day` and describe any possible problems with this model. (5 pts)

e) Fit a logistic regression model with survival as a function of `day` and `day^2` and compute the R^2 value. Plot the model residuals against `day` and describe any possible problems with this model. (4 pts)

f) Fit a logistic regression model with survival as a function of `temp` and `temp^2` and compute the R^2 value. Plot the model residuals against `day` and describe any possible problems with this model. (4 pts)

g) Fit a standard linear regression model to `temp` as a function of `day` and extract the residuals from this model. These residuals give an indication of whether a particular day of May was warmer or colder than average. Plot these residuals against survival and describe any patterns you see. (4 pts)

h) Fit a logistic regression model with survival as a function of the residuals from (g) and compute the R^2 value. Plot the model residuals against `day` and describe any possible problems with this model. (4 pts)

i) Create a table showing the ΔAIC values and Akaike weights for each of the four logistic regression models you fit above. Which model has the greatest support from the data? How do the other models compare to it? (4 pts)

j) Based on the results from (i), compute the model-averaged prediction of survival across all four models. Plot survival versus day and overlay your model-averaged prediction. Does this seem like good model overall? (5 pts)