Examining model diagnostics

QERM 514 - Homework 3

17 April 2020

Background

Section 7 of the U.S. Endangered Species Act (ESA) regulates situations in which a federal agency funds, permits, or otherwise has a "federal nexus" on any project that may influence a protected species. Federal agencies must seek a "consultation" on the project with either the U.S. Fish and Wildlife Service (USFWS) or the National Marine Fisheries Service (NMFS), depending on the species, and USFWS or NMFS must assure that any project does not cause "jeopardy" (a relatively high legal standard) for a protected species. A major conservation value of Section 7 consultation is the opportunity for USFWS and NMFS biologists to negotiate changes to projects that could minimize any negative impacts on species (or maximize any positive benefits).

The USFWS office in Lacey, Washington wanted to identify the characteristics of projects that would make them worthwhile for focused consultation time, with an emphasis on projects potentially impacting ESA-listed bull trout (*Salvelinus confluentus*). Experts developed assessments of the potential improvement(s) in a project that could be realized from negotiating changes to projects such as nearshore construction, culvert improvements, and riparian restoration. These assessment generated a unitless score of the potential value for 38 projects.

At this point the USFWS would like your assistance in evaluating a statistical model they hope to use for prioritizing project consultations. The accompanying data file `usfws_bull_trout.csv` contains 9 columns of information. They are

1. `score`: a project's potential value (numerical score on a scale of 0-15)
2. `stage`: 1 of 3 life history stage(s) occurring in the project area
    - adults (`A`)
    - juveniles/adults (`JA`)
    - eggs/juveniles (`EJ`)
3. `form`: 1 of 2 life history form(s) occurring in the project area
    - anadromous (`An`)
    - fluvial/anadromous (`FlAn`)
4. `cond`: 1 of 3 habitat conditions in the project area
    - pristine (`P`)
    - degraded (`D`)
    - highly degraded (`H`)
5. `risk`: 1 of 4 levels of extinction risk of the core population occurring in the project area
    - outside core area (`OC`)
    - low (`L`)
    - medium (`M`)

- high (`H`)

6. `unit`: 1 of 4 habitat unit types in the project area
    - inside a core area (`IC`)
    - outside a core area in freshwater (`OF`)
    - marine (`M`)
    - other (`OT`)
7. `prog`: whether or not the set of detailed management guidelines for projects of that type have been established
    - `Yes`
    - `No`
8. `BMP`: whether or not established best management practices will be followed in the project
    - `Yes`
    - `No`
9. `degflex`: the degree of flexibility in project design, timing, and location
    - low (`L`)
    - medium (`M`)

As you work through the following problems, make sure to explain your thought process and show all of your **R** code, so Mark can give you partial credit, if necessary.

Problems

a) Fit a linear model to the dataset that includes all 8 predictor variables. What is the R^2 for the model? Does is seem like a promising model?

b) Make a plot of the residuals against the model predictions. Name at least two things you should be looking for in a plot like this. What do you see?

c) Make a plot of the residuals against the predictor variable `stage`. Do you find this plot useful? Why or why not?

d) Produce a Q-Q plot of the model residuals and include a Q-Q line. Describe what you would hope to see here. Do you?

e) Would it make sense to plot e_t against e_{t+1} for this model? Explain why or why not.

f) Which projects have the 3 largest leverages? Briefly explain what this tells us.

g) What rule of thumb could you use to assess whether any leverages are particularly large? Under this rule of thumb, do you have any particularly large leverages? If yes, which projects?

h) Calculate the studentized residuals to look for outliers. Remember to use a Bonferroni correction, and explain why you should use it. What did you find? Which project has the largest studentized residual?

i) Calculate Cook's Distances and produce a halfnormal plot of them. Label the 3 largest D_i in the plot with the project names. Are these the same sites as the top 3 projects you identified in (g)? Briefly explain why or why not.

j) In a few sentences, summarize what you have learned about this analysis in terms of heteroscedasticity, normality, and influential observations.