Model Selection for Explanatory Models

Chapter 7, Lab 5

OpenIntro Biostatistics

Topics

- Building explanatory models
- Transforming variables
- Model comparison with adjusted R^2

In previous labs, multiple regression modeling was shown in the context of estimating an association while adjusting for possible confounders. This lab introduces explanatory modeling, in which the goal is to construct a model that explains the observed variation in the response variable. Explanatory modeling is concerned with identifying predictors associated with the response; there is no pre-specified primary predictor of interest.

The material in this lab corresponds to Section 7.8 in *OpenIntro Biostatistics*.

Introduction

Approaches to model selection vary from those based on careful study of a relatively small set of predictors to purely algorithmic methods that screen a large set of predictors and choose a final model by optimizing a numerical criterion. This course discusses model selection in the context of a small set of potential predictors.

Model selection for explanatory modeling follows these general steps:

1. *Data exploration*. Examine both the distributions of individual variables and the relationships between variables.
2. *Initial model fitting*. Fit an initial model with the predictors that seem most highly associated with the response variable, based on the data exploration.
3. *Model comparison*. Work towards a model with the highest adjusted R^2.
4. *Model assessment*. Use residual plots to assess the fit of the final model.

The process behind model selection will be illustrated with a case study in which a regression model is built to examine the association between the abundance of forest birds in a habitat patch and features of a patch.

Background Information

Habitat fragmentation is the process by which a habitat in a large contiguous space is divided into smaller, isolated pieces. Smaller patches of habitat are only able to support limited populations of organisms, which reduces genetic diversity and overall population fitness. Ecologists study habitat fragmentation to understand its effect on species abundance.

The forest.birds dataset in the oibiostat package contains a subset of the variables from a 1987 study analyzing the effect of habitat fragmentation on bird abundance in the Latrobe Valley of southeastern Victoria, Australia.[1]

The dataset consists of the following variables, measured for each of the 57 patches.

- abundance: average number of forest birds observed in the patch, as calculated from several independent 20-minute counting sessions.
- patch.area: patch area, measured in hectares. 1 hectare is 10,000 square meters and approximately 2.47 acres.
- dist.nearest: distance to the nearest patch, measured in kilometers.
- dist.larger: distance to the nearest patch larger than the current patch, measured in kilometers.
- altitude: patch altitude, measured in meters above sea level.
- grazing.intensity: extent of livestock grazing, recorded as either “light”, “less than average”, “average”, “moderately heavy”, or “heavy”.
- year.of.isolation: year in which the patch became isolated due to habitat fragmentation.
- yrs.isolation: number of years since patch became isolated due to habitat fragmentation.[2]

Data exploration

1.

Initial model fitting

Model comparison

Model assessment

[1] Loyn, R.H. 1987. "Effects of patch area and habitat on bird abundances, species numbers and tree health in fragmented Victorian forests." Printed in Nature Conservation: The Role of Remnants of Native Vegetation. Saunders DA, Arnold GW, Burbridge AA, and Hopkins AJM eds. Surrey Beatty and Sons, Chipping Norton, NSW, 65-77, 1987.

[2] The Loyn study completed data collection in 1983; yrs.isolation = 1983 − year.of.isolation.