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. Identify the variables in the dataset relevant for modeling the relationship between species abundance and features of a habitat; that is, the response variable and the potential predictor variables.
2. Explore the distribution of each variable with numerical and graphical summaries.
   a) Briefly describe the distribution of each variable.
   b) A common technique to improve model fit in linear regression (particularly in regards to achieving approximate linearity) is to transform variables that exhibit skew. A natural log transformation can help induce symmetry in right-skewed variables.

      Identify which variables could benefit from a natural log transformation. Apply the transformation and use the transformed version going forward.
   c) Examine the relationships between the predictor and response variables, as well as the relationships between predictor variables.
      i. Run the code in the template to create a scatterplot matrix. Each subplot in the matrix is a simple scatterplot; the variable names are listed along the diagonal of the matrix and the diagonal divides the matrix into symmetric plots.

[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`.

Describe what you see. Which variables seem to be strongly associated with the response? Do any predictor variables seem strongly associated with each other?

ii. Run the code in the template to create a correlation matrix. Confirm that the numerical summaries cohere with what you observed from the graphical summaries.

Initial model fitting

3. Based on the data exploration, which predictor variables should be included in an initial model?
4. Fit the initial model.
   a) Report the R^2 and adjusted R^2 of the model.
   b) Identify which variables are statistically significant at the $\alpha = 0.05$ level.

Model comparison

5. Fit models excluding the predictors that were not statistically significant. Based on comparing the adjusted R^2 values, consider whether any of these models are an improvement from the initial model.
6. The working model contains the grazing intensity variable. Only one of the coefficients associated with grazing intensity is statistically significant: heavy grazing. Individual categories of a categorical variable cannot simply be dropped, so a data analyst has the choice of leaving the variable as is, or collapsing it into fewer categories.

   For this model, it might be useful to collapse grazing-intensity into a two-level variable, with one category corresponding to the original classification of heavy, and another category corresponding to the other four categories.

   a) Create a plot of abundance versus grazing intensity. Does it seem that the distribution of abundance within the lowest four grazing intensity categories is roughly similar, relative to that within the highest category?
   b) Run the code in the template to create `grazing.binary`, which has levels `NotHeavy` and `Heavy`.
   c) Fit a model with the binary version of grazing intensity. Is this model an improvement over the model with the original version of grazing intensity?
7. Check whether incorporating an interaction term improves the model.
8. Report the variables in the final model and the model R^2.

Model assessment

9. Assess whether the residuals are normally distributed.

10. Run the code in the template to generate three plots that allow for a closer look at the residuals: a plot of residuals versus predicted abundance, and plots of residuals versus the two predictors.

    a) Recall that the definition of a residual is $e_i = y_i - \hat{y}_i$. Residual values closer to 0 are indicative of a more accurate prediction. In terms of comparing an observed value and a value predicted from a model, what does a large positive residual indicate? What does a large negative residual indicate?

    b) Examine the left and middle plot. For what predicted values of bird abundance do large positive residuals tend to occur, versus large negative residuals? For what values of area do large positive residuals versus large negative residuals tend to occur?

    c) In the middle plot, patches with heavy grazing are represented with red points. From the middle plot and right plot, assess how prediction error varies between patches where grazing intensity was between "light" and "moderately heavy" versus patches where grazing intensity was heavy.

Conclusions

11. Summarize the final model; interpret the model coefficients and R^2 value.

12. Ecologists might be interested in using the model to predict bird abundance based on features of a forest patch. Summarize the model accuracy, in terms accessible to a non-statistician.