Stat 428/528: Advanced Data Analysis 2

Chapters 9: Discussion of Response Models with Factors and Predictor

April 1, 2019

THE UNIVERSITY OF NEW MEXICO

Summary of the models

1. "Pure regression" model
   ——-observational or experimental data,
   ——response variable is continuous,
   ——only continuous predictor variables are considered,
   ——we are interested in the relationship between response and predictor variables.
   Example: Growth hormone is used as a prescription drug in medicine to treat children's growth disorders. In the medical study of short children, clinicians want to use the statistical relation to predict growth hormone deficiencies in short children by using simple measurements such as age and various body measurements of the children.

2. Pure ANOVA model
   ——-experimental data
   ———response variable is continuous,
   ———only qualitative factors are considered,
   ——-we are interested in the differences between the group means either at factor level or at combined factor levels
   Example: in the experiment comparing mean survival time of beetles, the potential effects of insecticide (with levels A, B, C, and D) and dose (with levels 1=low, 2=medium, and 3=high) are included in the model as factors.

3. ANCOVA or general model

——-designed experiments or observational studies

——-response variable is continuous,

——-both continuous predictors and qualitative factors are considered,

——-response variable is modeled as a linear combination of effects due to factors and predictors,

——-compare two (or more) groups while adjusting for one or more quantitative covariates.

——or comparing the relationship between the quantitative variables while accounting for group differences.

Example: Growth hormone is used as a prescription drug in medicine to treat children's growth disorders. In the medical study of short children, clinicians want to use the statistical relation to predict growth hormone deficiencies in short children by using simple measurements such as gender, age and various body measurements of the children.

Importance of “factor” statement

factor() statement defines which variables in the model are treated as factors.

- Each effect of Factor data type is treated as a factor.
- Effects in the model statement that are numeric data types are treated as predictors.
- To treat a measurement variable as a factor (with one level for each distinct observed value of the variable) instead of a predictor, convert that variable type to a factor using factor()

Example continued: in the experiment comparing mean survival time of beetles,

- insecticide (with levels A, B, C, and D)
- dose (with levels 1=low, 2=medium, and 3=high)
- now let’s reorder the dose given to each beetle on a measurement scale, such as 10, 20, and 30,

  ——- then the dosages can be used to define a predictor variable which can be used as a “regression effect in the model.

a. The simple additive model, or ANCOVA model, assumes that there is a linear relationship between mean survival time and dose, with different intercepts for the four insecticides.

   ```
   beetles$insect <- factor(beetles$insect)
   lm.t.i.d <- lm(times ˜ insect + dose, data = beetles)
   ```

b. A general model that allows separate regression lines for each insecticide is specified as follows:

   ```
   beetles$insect <- factor(beetles$insect)
   lm.t.i.d.id <- lm(times ˜ insect + dose + insect:dose,
   data = beetles)
   ```

c. two-way ANOVA model without interaction
   ——both dose and insecticide are treated as factors

```
beetles$insect <- factor(beetles$insect)
beetles$dose<- factor(beetles$dose)
lm.t.i.d <- lm(times ~ insect + dose, data = beetles)
```

d. two-way ANOVA model with interaction

```
lm.t.i.d.id <- lm(times ~ insect + dose + insect:dose,
data = beetles)
```

Model selection and case deletion

- Outliers tend to be cases with large residuals
  —-eliminating the largest residuals obviously makes the SSE and MSE smaller
- Variable selection methods tend to identify as good reduced models those with small MSEs
  —-Delete outliers if they are from recording errors (such as obvious typos), experimental accident (drop the tube) etc,.
  —-Usually after deleting outliers, new data will produce new outliers

Both variable selection and case deletion

- Cause the resulting model to appear better than it probably should
- Tend to give MSEs that are unrealistically small
- Prediction intervals are unrealistically narrow and test statistics are unrealistically large
- Test performed after variable selection or outlier deletion should be viewed as the greatest reasonable evidence against the null hypothesis, with the understanding that more appropriate tests would probably display a lower level of significance.

Final choice of a model based on:

- p-values, residual plots, other diagnostics
- Parsimony (Occam's Razor):
  ——the simplest plausible model among all reasonable models, given the data, is preferred.
- The sniff (giggle) test: does the model agree with expectations or theory? Do the signs make sense? Can you explain the results?
  ——-"statistically important v.s. "scientifically important
- Model validation studies
  ——Madigan and Raftery, in the 1994 edition of The Journal of the American Statistical Association, comment that "Science is an iterative process in which competing models of reality are compared on the basis of how well they predict what is observed; models that predict much less well than their competitors are discarded.

Hierarchy principle

Statisticians often follow the hierarchy principle, which states that

- a lower order term (be it a factor or a predictor) may be considered for exclusion from a model only if no higher order effects that include the term are present in the model.
  ——For example, given an initial model with effects A, B, C, and the A B interaction, the only candidates for omission at the first step are C and A B.
  ——If after C dropped, if A*B is significant, A is significant, but B is not significant, we will still keep A, B, and A*B in the model
- Note that a non-hierarchical backward elimination algorithm where single degree-of- freedom effects are eliminated independently of the other effects in the model is implemented in the step() procedure.