STAT425 Homework #5
Fall 2015 Nov 22, 2015

Homework 5

Due: Dec 03, 2015

Problem 1. (2+2+2=6pt) Analyze `warpbreaks` data as a two-way ANOVA.

(a) You are suggested to log-transformation on the response. Use the Box-Cox method to justify this suggestion.

(b) Determine which factors (including interactions) are significant.

(c) Now form a six-level factor from all combinations of the `wool` and `tension` factors. Which combinations—there are totally 15 pairs—are significantly different?

Problem 2. (4+6+4+6=20pt) The `barley` data can be found in the `lattice` package.

(a) Provide a graphical display of the data (see `R` code posted on Compass for HW5). The plot suggests the Morris data switched the years 1931 and 1932. So we will carry our analysis on a new data set called `newbarley`.

```
newbarley=barley
newbarley$year[newbarley$site=="Morris"]=
   ifelse(newbarley$year[newbarley$site=="Morris"]==1931, 1932, 1931)
```

Provide a graphical display of the new data.

(b) Perform a three-way ANOVA with `yield` as the response. Include all the two-way interactions, but no three-way interactions since we have only one observation in each three-way combination. Provide the ANOVA table, and comment on your result:

– are all the two-way interactions significant?

– If any, which two-way interaction(s) is(are) significant?

(c) Check the diagnostics (include a copy of the plots) — you will find that two points, the 23rd and the 83rd samples, stick out. They are the two with the highest residuals (in absolute value). Check these two points. Which site or sites are they from? Which year or years are they from?

(d) We suspect these two cases, the 23rd and the 83rd samples, were also switched. Can you find evidence from the 2nd graphical display produced in (a)? Switch the two cases, repeat the analysis: provide the ANOVA table, and comment on whether your result is similar to or different from the result you obtained in (b).

Problem 3. The `eggprod` comes from a randomized block experiment to determine factors affecting egg production.

a) (2pt) What is the blocking variable? (To answer this question, you need to read the description of this data set in R).

b) (6pt) Fit an additive two-way ANOVA model, and test whether there is any difference among the treatments. What's the value of your test statistic? What's its distribution under the null? What's the corresponding p-value?

c) (6pt) Fit a one-way ANOVA model (with only "`treat`" as the categorical predictor), and test whether there is any difference among the treatments. What's the value of your test statistic? What's its distribution under the null? What's the corresponding p-value?

d) (4pt) Recall that an F-test statistic takes the following form

$$F = \frac{(\text{Sum Sq})_1/m_1}{(\text{Sum Sq})_2/m_2},$$

which follows F_{m_1,m_2} under the null. Especially, the denominator $(\text{Sum Sq})_2/m_2$ can be viewed as an estimate of the error variance σ^2. You'll find that the numerators for the F-test statistics in (b) and (c) are the same, but the denominators are different. That is, (b) and (c) use different estimates for the error variance σ^2, denoted by $\hat{\sigma}_b^2$ and $\hat{\sigma}_c^2$, respectively.

What's the value for $\hat{\sigma}_b^2$, and what's the value for $\hat{\sigma}_c^2$?

Report the ratio, $\hat{\sigma}_c^2/\hat{\sigma}_b^2$, which is called the efficiency gained by the blocked design.

e) (2pt) Continue with the two-way additive model, i.e., the model we fit at (b). Which pairs of treatments are significantly different? (*Hint:* compute the Tukey HSD bands for all pairwise differences.)

Problem 4. (8pt) This problem is related to the Boston Housing data. You can download the data, "`BostonHousing.Rdata`", from the course website. The data has 506 observations on 16 variables.

Y	median value of owner-occupied homes in USD 1000's
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
lon	longitude of census tract
lat	latitude of census tract
crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft
indus	proportion of non-retail business acres per town
nox	nitric oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted distances to five Boston employment centres
rad	index of accessibility to radial highways
tax	full-value property-tax rate per USD 10,000
ptratio	pupil-teacher ratio by town
b	$1000(B - 0.63)^2$ where B is the proportion of blacks by town
lstat	percentage of lower status of the population

Implement the following variable selection methods to determine the "best" linear model with Y as the response and the other 15 variables as predictors:

- AIC
- AIC with stepwise
- BIC
- BIC with stepwise

Report the models selected by the four methods, and comment on your results: do they return the same model? If not, which method returns the smallest model, and which returns the largest model?