

## STAT 506, Spring 2017: Homework 4

- **Two factor data analysis.** A research laboratory was developing a new compound for the relief of severe allergies. In an experiment with  $N = 36$  subjects, the amounts of two active ingredients in the compound, factors A and B, were varied at three levels each. Randomization was used in assigning  $n = 4$  subjects to each of the  $3 \times 3 = 9$  treatment combinations. Commands you can swipe and paste into R follow; the response is hours of relief.

```
library(cfcdae); library(car); library(lsmeans)
hours=c(2.4,2.7,2.3,2.5,4.6,4.2,4.9,4.7,4.8,4.5,4.4,4.6,5.8,5.2,5.5,5.3,8.9,9.1,8.7,9.0,
  9.1,9.3,8.7,9.4,6.1,5.7,5.9,6.2,9.9,10.5,10.6,10.1,13.5,13.0,13.3,13.2)
a=factor(rep(1:3,each=12))
b=factor(rep(rep(1:3,each=4),3))
d=data.frame(hours,a,b)
levels(d$a)=c("low","medium","high")
levels(d$b)=c("low","medium","high")
```

- (a) Obtain an interaction plot for these data with 95% CI; have levels of factor B along the x-axis and levels of A connected. Does there appear to be an interaction between A and B? Sample code: `with(d,interactplot(b,a,hours,confidence=0.95))`.
  - (b) Fit the interaction model and test  $H_0 : \alpha\beta_{ij} = 0$  at the 5% level. What do you conclude? Sample code: `f=lm(hours~a*b)`, followed by `Anova(f,type=3)`.
  - (c) Obtain pairwise differences in A sliced at levels of B. Use the `lsmeans` package. Describe the significant differences in A at each level of B. Sample code `pairs(lsmeans(f,"a",by="b"))` and `confint(pairs(lsmeans(f,"a",by="b")))`.
  - (d) Two factors have a *reinforcement* or *synergistic* interaction if the response is *greater* than what we would expect under an additive model. Is that what we have here with the two allergy medication ingredients? What levels of A and B provide the greatest amount of relief?
  - (e) Obtain the standard R diagnostic plot; comment on the constant variance and normality assumptions. Sample code: `par(mfrow=c(2,2))` followed by `plot(f)`.
- **Three factor data analysis.** Assemblers in an electronics firm attach 12 components to a newly developed “board” that will be used in automatic-control equipment in manufacturing plants. An analyst conducted an experiment to study the effects of three factors on the mean time to assemble a board. Factor A is gender ( $i = 1, 2$  for male/female), B is the sequence of assembling the components ( $j = 1, 2, 3$ ), and C was the amount of experience the assemblers had ( $k = 1$  for under 1.5 years and  $k = 2$  for over 1.5 years). Note that the only treatment here is the sequence B; gender and experience are effectively blocked on. The experiment was set up so that within each pairing of gender and experience, 15 individuals were randomly assigned the three sequences, i.e. there are  $n = 5$  replications of each gender/experience/treatment combination. The total time in minutes to assemble 50 boards was observed for each subject. The following R code reads in the data:

```
d=read.table("http://people.stat.sc.edu/hansont/stat506/electronics.txt",header=F)
colnames(d)=c("time","gender","sequence","experience","rep")
d$gender=factor(d$gender)
d$sequence=factor(d$sequence)
d$experience=factor(d$experience)
levels(d$gender)=c("male","female")
levels(d$sequence)=c("1","2","3")
levels(d$experience)=c("<18 months",">=18 months")
```

- (a) Fit a full interaction model with the main-effects, all two-way interactions, and the three-way interaction, e.g. using `f1=lm(time~gender*sequence*experience,data=d)`, followed up by Type III tests for dropping terms, e.g. `Anova(f1,type=3)`. The `Anova` function is in the `car` package. Note which p-values are larger than 0.05 and carry out one overall F-test that you can drop these effects, e.g. using `anova(f2,f1)`.
  - (b) Using the *reduced* model from part (a), find which sequence(s) are significantly best (i.e. take the least amount of time).
  - (c) Look at R's standard residual plots and comment on overall model fit.
  - (d) Are the experience and gender factors significant? Who takes less time to assemble boards, males or females? How about more vs. less experienced?
- **Randomized complete block design analysis.** A researcher studied the effects of three experimental diets with varying fat contents on the total lipid (fat) level in blood plasma. Total lipid level is a widely-used predictor of coronary heart disease. Fifteen male subjects who were within 20% of their ideal body weight were grouped into five blocks according to their age. Within each block the three experimental diets were randomly assigned to the three subjects. Data on the *reduction* in lipid level (grams/liter) after the subjects were on the diet for a fixed period of time follow.

```
reduction=c(0.73,0.86,0.94,1.40,1.62,0.67,0.75,0.81,1.32,1.41,0.15,0.21,0.26,0.75,0.78)
fat=factor(c(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3))
age=factor(c(1,2,3,4,5,1,2,3,4,5,1,2,3,4,5))
d=data.frame(reduction,fat,age)
levels(d$age)=c("15-24 yrs","25-34 yrs","35-44 yrs","45-54 yrs","55-64 yrs")
levels(d$fat)=c("extremely low","fairly low","moderately low")
```

- (a) Obtain an interaction plot (also called a “spaghetti plot” or “profile plot”) as in the notes where the age category is along the x-axis and the three diet types are connected. Is the plot approximately parallel? Which diet provides the *greatest reduction* in lipids?
- (b) Perform Tukey's test for additivity as shown in the notes. What do you conclude? Is the additive model okay?
- (c) Fit the additive model to these data followed by `Anova(f,type=3)`. Is the treatment (fat in diet) significant? Perform pairwise comparisons across diets using Tukey's HSD. Do the diets provide significantly different lipid reduction? Provide a “lines” plot showing significant differences.
- (d) Obtain R's standard diagnostic plots and comment on whether modeling assumptions (constant variance, normality) are reasonable.