STAT 506, Spring 2017: Homework 5

- **Nested design**. A large manufacturing company operates three regional training schools for mechanics, in each of its operating districts Atlanta, Chicago, and San Francisco. The schools have two instructors each, who teach classes of about 15 mechanics in three-week sessions. The company was concerned about the effect of school (factor A) and instructor (factor B) on the learning achieved. To investigate these effects, classes in each district were formed in the usual way and then randomly assigned to one of the two instructors in the school. A summary mesure of the amount of learning in each class was recorded (higher is better).

```
library(cfcdae); library(car); library(lsmeans)
learning=c(25,29,14,11,11,6,22,18,17,20,5,2)
school=factor(c(rep("Atlanta",4),rep("Chicago",4),rep("San Francisco",4)))
instructor=factor(c(1,1,2,2,1,1,2,2,1,1,2,2))
f=lm(learning~school+school/instructor)
Anova(f,type=3)
lsmeans(f,"school") # mean learning averaged over instructors
pairs(lsmeans(f,"school")) # which school(s) best?
```

  (a) Fit the nested model $y_{ij} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ij}$. Report the p-values for testing $H_0 : \alpha_i = 0$ and $H_0 : \beta_{j(i)} = 0$. Are schools significantly different in how much learning is achieved? How about instructors?

  (b) Which school is best? Draw a lines plot indicating significant differences among the three schools.

  (c) Obtain R's standard diagnostic plot and comment on constant variance and normality.

- **ANCOVA**. In an experiment to investigate the effect of paper color (blue, green, orange) on response rates for questionaires distributed by the "windshield method" in supermarket parking lots, $N = 15$ lots were chosen in a metropolitan area and each color assigned at random to five of the lots; the response rate as a percentage was recorded. The size of the lots was also recorded as a concomitant variable that might affect the percentage response rate.

```
percent=c(28,26,31,27,35,34,29,25,31,29,31,25,27,29,28)
color=factor(c(rep("blue",5),rep("green",5),rep("orange",5)))
spaces=c(300,381,226,350,100,153,334,473,264,325,144,359,296,243,252)
plot(percent~spaces,pch=19,col=c("blue","green","orange")[color])
```

  (a) Based on the scatterplot, describe what happens to the response percent as the number of parking spaces increases. Which color appears to be best? Does the ANCOVA model appear to be adequate?

  (b) Fit the model ignoring the number of spaces, i.e. $y_{ij} = \mu + \tau_i + \epsilon_{ij}$. Here $i = 1, 2, 3$ for blue, green, & orange. Is color significant?

  (c) Now include the number of spaces as a concomitant variable, i.e. $y_{ij} = \mu + \tau_i + \gamma x_{ij} + \epsilon_{ij}$. Is color significant when we adjust for the size of the parking lot?

  (d) Obtain pairwise comparisons via Tukey and make a lines plot. Which color(s) is/are significantly best? By how much?

  (e) Obtain R's standard diagnostic plot and comment on constant variance and normality.

- **Repeated measures within one factor**. A repeated measures study was conducted to examine the effects of two different store displays for a household product (factor A) on sales in four successive time periods (factor B). Eight stores were randomly selected and four assigned at random to each display. The raw sales were recorded in each time period, i.e. sales are repeated over time but not display.

```
library(lme4); library(RLRsim)
sales=c(956,953,938,1049,1008,1032,1025,1123,350,352,338,438,412,449,385,532,
        769,766,739,859,880,875,860,915,176,185,168,280,209,223,217,301)
display=factor(c(rep(1,16),rep(2,16)))
time=factor(rep(1:4,8))
store=factor(rep(1:8,each=4))
d=data.frame(sales,display,time,store)
with(d,interactplot(time,display,sales,confidence=0.95)) # parallel?
f=lmer(sales~display*time+(1|store))
Anova(f,type=3)
exactRLRT(f)

f=lmer(sales~display+time+(1|store)) # additive model
Anova(f,type=3) # display significant?  time significant?
exactRLRT(f) # tests H0: sigma_rho=0
plot(f) # residuals vs. fitted
qqnorm((ranef(f)$store[,1])) # NPP of estimated rho_i
```

(a) What does the interaction plot tell you? Does there appear to be a signficant difference across displays? Is additivity between display and time reasonable?

(b) Fit the repeated measures model $y_{ijk} = \mu + \rho_i + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}$ where $i = 1, \ldots, 8$ denotes store, $j = 1, 2$ is the display, and $k = 1, 2, 3, 4$ is the time period. Here $\rho_1, \ldots, \rho_8 \overset{iid}{\sim} N(0, \sigma_\rho^2)$ are random store effects. Is the `display:time` interaction significant? If not, refit the same model without the interaction.

(c) Does the display type significantly affect sales? How about the time period?

(d) For significant factors in (b), obtain a lines plot using Tukey's HSD.

(e) Test $H_0 : \sigma_\rho^2 = 0$ vs. $H_a : \sigma_\rho^2 > 0$.

(f) For your fitted model examine a plot of the residuals vs. fitted and comment on constant variance. Comment on the normal probability plot of the $\hat{\rho}_i$. Is the normality assumption on the random effects valid?