Analysis of Covariance

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 506: Introduction to Experimental Design

ANalysis of **COVA**riance

Add a continuous predictor to an ANOVA model = ANCOVA.

- Mix continuous and discrete predictors.
- Useful for testing treatment effects in presence of continuous predictor(s) that may explain much variability.
- Continuous predictor may be concomitant (supplemental, uncontrolled) or controlled (e.g. drug dose in *mg*).
- Concomitant variable should be unaffected by treatments; i.e. they should be "independent." They are often measured before study takes place.
- Examples: prestudy attitude, age, SES, aptitude, baseline outcomes (e.g. seizure rate).
- Often the same types of variables one might block on in a RCBD.

Simplest ANCOVA model

One treatment and one covariate that enters model linearly. Have $i = 1, \ldots, r$ treatment levels and $j = 1, \ldots, n_i$ observations within level i. Model is

$$y_{ij} = \mu + \tau_i + \gamma x_{ij} + \epsilon_{ij}.$$

This gives r parallel regression lines, one for each treatment level (a picture helps). Fixing x, the mean difference between group i and group j is

$$\mu + \tau_i + \gamma x - (\mu + \tau_j + \gamma x) = \tau_i - \tau_j.$$

Can get from `lsmeans`, `pairwise`, etc.

ANOVA table

For the simple ANCOVA model, the ANOVA table will have a row for the concomitant variable and another row for the treatment effects.

The p-values test $H_0 : \gamma = 0$ (concomitant variable not important) and $H_0 : \tau_1 = \cdots = \tau_r = 0$ (no treatment differences).

Cracker sales

- CRD where $N = 15$ stores were randomly assigned one of three "promotion" treatment levels:
  1. $i = 1$ sampling of product by customers in store and regular shelf space,
  2. $i = 2$ additional shelf space,
  3. $i = 3$ special display shelves at ends of aisle in addition to regular shelf space.
- y_{ij} is number of cases sold during the promotional period.
- x_{ij} is number of cases sold during the previous (non-promotional) period.
- Model fit in R is $y_{ij} = \mu + \tau_i + \gamma x_{ij} + \epsilon_{ij}$.

Cracker sales in SAS

```
library(cfcdae); library(lsmeans); library(car)
treatment=factor(c(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3))
cases    =c(38,39,36,45,33,43,38,38,27,34,24,32,31,21,28)
preceding=c(21,26,22,28,19,34,26,29,18,25,23,29,30,16,29)
d=data.frame(cases,preceding,treatment)

plot(cases~preceding,pch=19,col=c("green","blue","red")[treatment])
legend(17,45,legend=c("1","2","3"),col=c("green","blue","red"),pch=19)

f=lm(cases~preceding+treatment)
Anova(f,type=3)
lsmeans(f,"treatment")
pairs(lsmeans(f,"treatment"))
pairwise(f,treatment)

library(HH) # has a nice function
ancova(cases~preceding+treatment,data=d)
```

Checking for non-constant slopes

The assumption of parallel slopes should be checked, via plots and/or Type III tests. A model that allows for slopes to change with treatment is

$$y_{ij} = [\mu + \tau_j] + [\gamma + \gamma_j]x_{ij} + \epsilon_{ij}.$$

```
f2=lm(cases~preceding*treatment)
Anova(f2,type=3) # p=0.4 so additive model okay
```

Diagnostics?

Generalizations

- Basic model is $y_{ij} = \mu + \tau_i + \gamma x_{ij} + \epsilon_{ij}$.
  - Response mean is linear function of x for each treatment group: parallel lines.
  - $i = 1, \ldots, r$ levels of one treatment modeled.
  - $\tau_i - \tau_j$ gives mean treatment differences for a given level of x.
  - Similar, but *simpler* than a RCBD with x chopped up into categories like age group. Just treat age as continuous.
  - Increased *efficiency* if age really is linear.
- Nonlinear mean, e.g. $y_{ij} = \mu + \tau_i + \gamma_1 x_{ij} + \gamma_2 x_{ij}^2 + \epsilon_{ij}$.
  - Mean response is parallel *curves* in x, one for each treatment level.
  - Might be necessary if e_{ij} vs $\hat{y}_{ij}$ shows a parabolic (or otherwise nonlinear) shape.
  - $\tau_i - \tau_j$ again gives mean treatment differences for a given level of x.

Generalizations

- More factors, e.g. $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma x_{ijk} + \epsilon_{ijk}$.
  - Here $i = 1, \ldots, a$ levels of A, $j = 1, \ldots, b$ levels of B, and $k = 1, \ldots, n_{ij}$ replicates in $A = i$ and $B = j$.
  - If this fits, should see approximately parallel curves in scatterplot stratified by (i, j).
  - If $H_0 : (\alpha\beta)_{ij} = 0$ then analysis simplifies; can look at differences in main effects. Pairwise difference, e.g. $\beta_3 - \beta_1$ do not change with either i or x.
- More concomitant variables, e.g. $y_{ijk} = \mu + \tau_i + \gamma_1 x_{i1k} + \gamma_2 x_{i2k} + \epsilon_{ijk}$ where x_{ijk} is variable j on kth subject with treatment i.
  - Mean response is parallel *surfaces* in (x_1, x_2).
  - Here we are assuming parallel *planes*, one for each level of i.

Salable flowers

- Factor A is flower variety: $i = 1$ LP, $i = 2$ WB.
- Factor B is moisture level: $j = 1$ low, $j = 2$ high.
- $N = 24$ plots total; $n_{ij} = 6$ replications of each pairing (i, j).
- y_{ijk} is number of flowers horticulturist can sell.
- x_{ijk} is plot size; expect $\gamma > 0$.
- Model is $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma x_{ijk} + \epsilon_{ijk}$.
- CRD with factorial treatment structure.

Salable flowers in SAS

```
variety= factor(c(1,1,1,1,1,1,2,2,2,2,2,2,1,1,1,1,1,1,2,2,2,2,2,2))
moisture=factor(c(1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2))
yield=   c(98,60,77,80,95,64,55,60,75,65,87,78,71,80,86,82,46,55,76,68,43,47,62,70)
plotsize=c(15, 4, 7, 9,14, 5, 4, 5, 8, 7,13,11,10,12,14,13, 2, 3,11,10, 2, 3, 7, 9)
d=data.frame(yield,plotsize,variety,moisture)
plot(yield~plotsize,col=rep(1:4,each=6),main="yield by plotsize & variety:moisture",pch=19)
legend(3,90,legend=c("1:1","2:1","1:2","2:2"),col=1:4,pch=19)

f1=lm(yield~plotsize+variety*moisture,data=d)
Anova(f,type=3)
f2=lm(yield~plotsize+variety+moisture,data=d)
pairs(lsmeans(f2,"variety"))
pairs(lsmeans(f2,"moisture"))
```

Let's look at diagnostics...