STA 2004F Homework 2 Solutions.

1. Two types of extended Latin square designs were given: (a) 3 replicates of a 4×4 design, and (b) intermixed 4×4 Latin squares. What types of systematic error are eliminated using design (a)? What types of systematic error are eliminated using design (b)?

   This question is taken from PE, Example 3.9. Quoting from that book: "In both designs constant differences between days have no effect on the treatment comparisons. In the second design, the effect of constant differences between times of day persisting throughout the whole experiment is likewise eliminated. In (a), however, not only is this done, but also time of day effects are eliminated separately from each set of four days. This would be particularly useful if, as might be convenient, there is a considerable gap in time between the sets of four days, or it it were desired to introduce some external change in conditions, either of which things might mean that time of day effects would not be the same in all parts of the experiment."

   Most people also noted that design (b) has more degrees of freedom for estimating σ^2, so if there was no variation among the three replicates design (b) could be more efficient.

2. *Analysis of covariance in the randomized block design*: Suppose we have observations in an RB design, and a baseline variable z measured for each experimental unit. The linear model extension of the CR design is

$$y_{js} = \mu + \tau_j + \beta_s + \gamma(z_{js} - \bar{z}_{..}) + \epsilon_{js}, \quad j = 1, \ldots, v; \quad s = 1, \ldots, r.$$

   (a) Derive the least squares estimates of μ, τ_j, β_s and γ, under the summation constraints $\sum \tau_j = 0, \sum \beta_s = 0$.
   Minimizing $\sum_{js}\{y_{js} - \mu - \tau_j - \beta_s - \gamma(z_{js} - \bar{z}_{..})\}^2$ over μ, τ_j, β_s and γ, invoking the summation constraints, leads to

$$\begin{aligned} \hat{\mu} &= \bar{y}_{..}, \\ \hat{\tau}_j &= \bar{y}_{j.} - \bar{y}_{..} - \hat{\gamma}(\bar{z}_{j.} - \bar{z}_{..}), \\ \hat{\beta}_s &= \bar{y}_{.s} - \bar{y}_{..} - \hat{\gamma}(\bar{z}_{.s} - \bar{z}_{..}), \\ \hat{\gamma} &= \frac{\sum_{js}(y_{js} - \bar{y}_{j.} - \bar{y}_{.s} + \bar{y}_{..})(z_{js} - \bar{z}_{j.})}{\sum_{js}(z_{js} - \bar{z}_{j.} - \bar{z}_{.s} + \bar{z}_{..})(z_{js} - \bar{z}_{j.})}, \end{aligned}$$

   the latter following directly upon substituting for $\hat{\tau}_j$, $\hat{\beta}_s$ and $\hat{\mu}$ in the equation for $\hat{\gamma}$. It can be verified that the expression for $\hat{\gamma}$ is equal to

$$\hat{\gamma} = R_{zy}/R_{zz}$$

   where $R_{zy} = \sum_{js}(y_{js} - \bar{y}_{j.} - \bar{y}_{.s} + \bar{y}..)(z_{js} - \bar{z}_{j.} - \bar{z}_{.s} + \bar{z}_{..})$ and R_{zz} is defined analogously. Note that if you are not careful it is easy to get the incorrect formula $R_{zy}/\sum_{js}(z_{js} - \bar{z}_{..})^2$ for $\hat{\gamma}$. This is wrong because the denominator SS is not correct. Note also that an easy way to compute R_{zz} is to fit a randomized block model with block and treatment effects to the **before** values; the residual SS from this fit is R_{zz}.

(b) Show that the mean square of the residuals is an unbiased estimate of σ^2, under the second moment conditions on the ϵ's.

I think now that the easiest way to get this is to first reduce the residual sum of squares to

$$\sum_{js} \{\epsilon_{js} - \bar{\epsilon}_{j.} - \bar{\epsilon}_{.s} + \bar{\epsilon}_{..} - \hat{\gamma}(z_{js} - \bar{z}_{j.} - \bar{z}_{.s} + \bar{z}_{..})\}^2,$$

either by substituting the formulae for $\hat{\mu}$, $\hat{\tau}_j$ and $\hat{\beta}_s$, or arguing that the answer obviously can't depend on μ, τ_j or β_s, so setting these values to zero. Now use the results $E(\hat{\gamma}) = 0$, $\text{var}(\hat{\gamma}) = \sigma^2/R_{zz}$ to complete the calculation. Note also that we already proved, in the context of the simple RB design, that $E\sum(\epsilon_{js} - \bar{\epsilon}_{j.} - \bar{\epsilon}_{.s} + \bar{\epsilon}_{..})^2 = (r-1)(v-1)$.

3. *Anocova continued*

(a) Use your favorite computer package to find the treatment means, adjusted for the covariate $z_{js} - \bar{z}_{..}$ (the first count).

For my answer I'm just going to give segments of code, even though you are not allowed to do that! The last two lines give the adjusted treatment means. The coefficient estimates computed in `R` do **not** seem to be the least squares estimates given in 2(a), even after imposing the summation constraints. If they were, then the command `coef(anocov)[3:10]+mean(after)` would give the adjusted means, but it doesn't. The treatment contrasts however are all identical. If I marked your answers incorrect and the problem was with `R`, please bring your homework by so I can correct the grades if needed.

```
> options(contrasts=c("contr.sum","contr.poly"))
> anova(anocov<-aov(after ~ before + tmt + block))
Analysis of Variance Table

Response: after
          Df Sum Sq Mean Sq F value  Pr(>F)
before     1 408441  408441   57.27 7.2e-09 ***
tmt        8 223465   27933    3.92  0.0022 **
block      3 110055   36685    5.14  0.0047 **
Residuals 35 249605    7132
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> coef(anocov)
(Intercept)      before        tmt1        tmt2        tmt3        tmt4
  82.859051    1.559010   90.825598  -82.018039   74.947860   26.960400
       tmt5        tmt6        tmt7        tmt8      block1      block2
 -13.385889 -105.586582    6.011466   81.777184  -76.412118   25.444044
     block3
  63.988283

 > adj_mean <- tapply(after,tmt,mean)-
 1.559*(tapply(before,tmt,mean)-mean(before))
 > adj_mean

       0      1CK      1CM      1CN      1CS      2CK      2CM      2CN
373.9525 201.1090 358.0748 310.0870 269.7408 177.5410 289.1385 364.9038
     2CS
203.5950
```

Show that the variance for comparing a treatment mean to the control treatment is given by

$$\sigma^2 \left\{ \frac{1}{r_1} + \frac{1}{r_2} + \frac{(\bar{z}_{1.} - \bar{z}_{2.})^2}{R_{zz}} \right\}$$

where r_2 is the number of observations on the treatment of interest, r_1 is the number of observations on the control, $\bar{z}_{2.}$ and $\bar{z}_{1.}$ are the means of the covariate on treatment and control, respectively, and R_{zz} is the residual sum of squares of the z_{js}'s, within treatments, after eliminating block effects. (Your estimate of γ obtained above should be R_{yz}/R_{zz}, analogously to that derived in class.)

$$\begin{aligned}\text{var}\{\bar{y}_{j.} - \bar{y}_{0.} - \hat{\gamma}(\bar{z}_{j.} - \bar{z}_{0.})\} &= \text{var}(\bar{y}_{j.}) + \text{var}(\bar{y}_{0.}) + (\bar{z}_{j.} - \bar{z}_{0.})^2 \text{var}(\hat{\gamma}) + \text{covariance terms} \\ &= \sigma^2/r_1 + \sigma^2/r_2 + \sigma^2(\bar{z}_{j.} - \bar{z}_{0.})^2/R_{zz},\end{aligned}$$

using the properties of $\hat{\gamma}$ above. The covariance terms are zero because $\hat{\gamma}$ and $\bar{y}_{j.}$ are uncorrelated for any j. Why? Because $\hat{\gamma}$ is constructed from the residuals $y_{js} - \bar{y}_{j.} - \bar{y}_{.s} + \bar{y}_{..}$ and they are orthogonal to $\bar{y}_{j.}$.

(b) Use the residual sum of squares after fitting the full model to estimate the variance in part (a). Which of the treatments applied gives a significant reduction in eelworm counts?
First compute R_{zz}:

```
> anova(aov(before~tmt+block))
Analysis of Variance Table

Response: before
          Df Sum Sq Mean Sq F value  Pr(>F)
tmt        8  29142    3643    1.08     0.4
block      3 159617   53206   15.78 1.0e-06 ***
Residuals 36 121409    3372

Rzz <- 121409
```

and then the variances for comparing each adjusted treatment mean to the control:

```
> zbar<-tapply(before,tmt,mean)
> zbar[-1]-zbar[1]
    1CK     1CM     1CN     1CS     2CK     2CM     2CN     2CS
 19.062   4.812 -22.938 -19.188  71.062  18.562 -26.188  15.062
```

```
adj_var<- (1/16 + 1/4 + .Last.value^2/Rzz)*7132
sqrt(adj_var)
     1CK      1CM      1CN      1CS      2CK      2CM      2CN      2CS
47.43518 47.22405 47.53585 47.43814 50.25334 47.42353 47.63439 47.35058
> adj_mean[-1]-adj_mean[1]
        1CK         1CM         1CN         1CS         2CK         2CM
-172.843437  -15.877687  -63.865437 -104.211687 -196.411438  -84.813937
        2CN         2CS
  -9.048687 -170.357437
```

These estimated contrasts with their estimated standard errors can be used to assess which of the treatments is most effective. This can be done either by significance tests or confidence intervals, and you can use either the normal critical values, or the t critical values with $(r-1)(v-1)-1$ degrees of freedom. If you like, you can make a multiple testing correction using a Bonferroni adjustment (setting $\alpha.05/8$, for example). Alternatively you can use the studentized range test given below. In any case we see that treatments 2CK and 1CK give the largest reduction in eelworm count, followed by 2CS and 1CS, and that these reductions are statistically significant.

4. Tukey's studentized range test

Problem In a randomized block design, let $u_j = \bar{y}_{j.} - (\mu + \tau_j)$. Show that $Eu_j = 0$, and var$u_j = \sigma^2/r$. Use the above result to show that

$$\Pr\{\frac{\max u_j - \min u_j}{(MS_{resid}/r)^{1/2}} \le q_{v,\nu,\alpha}\} = 1 - \alpha$$

where $q_{v,\nu,\alpha}$ is the $1-\alpha$ critical value for the studentized range distribution. Deduce that

$$Pr\{|u_j - u_{j'}| \le \sqrt{\frac{MS_{resid}}{r}}\; q_{k,\nu,\alpha} \text{ for all } j, j'\} = 1 - \alpha.$$

Use this to show that a set of simultaneous $100\alpha\%$ confidence intervals for all pairwise treatment differences $\tau_j - \tau_{j'}, j \ne j'$, is given by

$$\{(\bar{y}_{j.} - \bar{y}_{j'.}) \pm q_{v,\nu,\alpha}\sqrt{MS_{resid}/r}\}.$$

Answer This question falls out easily, once X_i is identified with u_j, R with $\max u_j - \min u_j$, and MS_{resid}/r with an unbiased estimate of var(u_j). Again, MS_{resid}/r is independent (under normality) of u_j because it is formed from the residuals. The simultaneous confidence limits are obtained by pivoting on the u_j.

5. *(CR 3.2): Optional for M.Sc.* Suppose in a matched pair design the responses are binary. Construct the randomization test for the null hypothesis of no treatment difference. Compare this with the test based on that for the binomial model, where Δ is the log odds ratio.