Technical note on analysis of covariance

This follows, but is slightly simpler than, CR pp. 59,60. Suppose we fit the following model with treatment effects and a single covariate:

$$y_{js} = \mu + \tau_j + \gamma(z_{js} - \bar{z}_{..}) + \epsilon_{js}$$

where τ_j is the treatment effect associated with the jth treatment, and z_{js} is the value of a covariate measured on the sth unit to receive treatment j. We make the usual second moment assumptions about ϵ_{js}. Our main interest is in comparing two treatments via the contrast $\tau_j - \tau_{j'}$, say.

First, note that

$$E\bar{y}_{j.} = \mu + \tau_j + \gamma(\bar{z}_{j.} - \bar{z}_{..})$$

so that

$$\hat{\mu} + \hat{\tau}_j = \bar{y}_{j.} - \hat{\gamma}(\bar{z}_{j.} - \bar{z}_{..})$$

and hence

$$\hat{\tau}_j - \hat{\tau}_{j'} = \bar{y}_{j.} - \bar{y}_{j'.} - \hat{\gamma}(\bar{z}_{j.} - \bar{z}_{j'.})$$

showing as expected that the difference in the unadjusted treatment means is not consistent for the true treatment effect under the assumed linear model.

We can also show that

$$\text{var}(\hat{\tau}_j - \hat{\tau}_{j'}) = \sigma^2 \left\{ \frac{2}{r} + \frac{(\bar{z}_{j.} - \bar{z}_{j'.})^2}{\sum_{js}(z_{js} - \bar{z}_{j.})^2} \right\},$$

using the results that

$$\begin{aligned} \hat{\gamma} &= \sum_{js}(y_{js} - \bar{y}_{j.})(z_{js} - \bar{z}_{j.}) / \sum_{js}(z_{js} - \bar{z}_{j.})^2 \\ \text{var}(\hat{\gamma}) &= \sigma^2 / \sum_{js}(z_{js} - \bar{z}_{j.})^2. \end{aligned}$$

Note that the variance is inflated relative to that of the difference of unadjusted means.

Now, if adjustment by the covariate z was unnecessary, this represents a loss of precision for the estimated comparison of the treatments. How much is this loss likely to be? One approach to this is to consider the randomization expectation of the variance inflation factor. Since treatments are assigned at random to units, the collection of observed z_{js} is a random permutation of the n possible values. Under this randomization distribution we have:

$$E_R \sum_{js}(z_{js} - \bar{z}_{j.})^2 = k\sigma_z^2$$

where k is the degrees of freedom ($v(r-1)-1$, I think) for the error sum of squares in the analysis of variance table, and σ_z^2 is the variance of the z_{js} in the population (i.e. $\sum_{js}(z_{js}-\bar{z}_{..})^2/n$.)

Also

$$E_R(\bar{z}_{j.}-\bar{z}_{j'.})^2 = 2\sigma_z^2/r$$

by the usual formula for the difference between two means. Putting these together we have

$$E_R\frac{(\bar{z}_{j.}-\bar{z}_{j'.})^2}{\sum_{js}(z_{js}-\bar{z}_{j.})^2} \doteq \frac{2}{rk}$$

so the inflation factor is approximately $1+1/k$, and if the residual degrees of freedom is large relative to the sample size, we can think of the cost of adding an unnecessary covariate as approximately equal to the loss of one observation.

This argument goes through for q covariates z, (where the loss is approximately q observations) and for a RB design, where in the latter case the formula for k given above would need to be changed.

The assumption that the covariate is uncorrelated with the response is needed for the derivation of $E_R\sum_{js}(z_{js}-\bar{z}_{j.})^2$ above to be correct. Also, if the covariate is correlated with the response and a model with no covariates is used, then the variance is inflated relative to the correct linear model.