Covariates

Gary W. Oehlert

School of Statistics
University of Minnesota

November 14, 2013

What is a Covariate?

In our context, a covariate is a predictive response. A predictive response is correlated with (predictive of) the primary response.

We cannot measure the predictive response ahead of time. This means that we cannot block on it.

However, we can use the covariate to “model away” some of the variance. This achieves variance reduction through modeling rather than blocking or similar.

The only real design aspect of a covariate is that you must plan to measure it.

Your experiment can be CRD or RCB or whatever, you just need to measure the covariate and then use it in the model.

Example

Construal level theory says that you tend to think about distant (space, time, etc) things in abstract/ general ways and close things in concrete/ detailed ways.

The experiment asks subjects to rate an advertisement. They are told they will be buying a camera either today or in one year. They are then shown an advertisement, which either talks about the LCD screen size (a secondary general feature) or the quality of the lens (a primary specific feature).

All subjects are also asked to rate the importance of the lens on a camera and the LCD screen on the camera.

People who know that the lens is important will likely react more favorably to the information about the lens, whether it is near or distant in time.

If we knew their feelings about lenses, we could have blocked on that. But we don't know their feelings about lenses until we run the experiment.

We cannot block on belief in "lens importance," but we can model out some of the variability using a covariate (or two!).

Covariate Models

For simplicity, assume we have a single factor treatment and a single covariate x. The basic covariate model is

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \epsilon_{ij}$$

for $i = 1, \ldots, g$ and $j = 1, \ldots, n_i$.

The covariate model looks like an ordinary fixed effects model with the addition of a regression-like term.

The model assumes a linear relationship, but assuming does not make it true.

As with any regression, we need to check for a linear relationship.

A transformation of the covariate or the response could improve linearity.

We can also consider higher polynomial terms in the covariate, but that is less common.

Family of covariate models

Single line (no treatment effects): $y_{ij} = \mu + \beta x_{ij} + \epsilon_{ij}$

Parallel lines or separate intercepts model (treatments affect the mean response or intercept but not the relationship with the covariate): $y_{ij} = \mu + \alpha_i + \beta x_{ij} + \epsilon_{ij}$ with $\sum \alpha_i = 0$

Single intercept model (treatments affect the relationship with the covariate but not the mean response or intercept): $y_{ij} = \mu + \beta x_{ij} + \beta_i x_{ij} + \epsilon_{ij}$ with $\sum \beta_i = 0$

Separate lines model (treatments affect slopes and intercepts): $y_{ij} = \mu + \alpha_i + \beta x_{ij} + \beta_i x_{ij} + \epsilon_{ij}$ with $\sum \beta_i = 0$ and $\sum \alpha_i = 0$

The single line model is a special case of the parallel lines model, which is in turn a special case of the separate lines model. We can compare these three via anova.

The single line model is a special case of the separate slopes model, which is in turn a special case of the separate lines model. We can compare these three via anova.

Parallel lines model and separate slopes model can be compared via AIC or BIC.

The separate slopes model depends delicately on how the covariate is centered. That is, reexpressing x_{ij} by adding 10 to all the values leads to a fundamentally different (and possibly worse or better fitting) separate slopes model.

Note: Sometimes things are easier to understand with a central value plus offset, and sometimes they're easier to understand if you just combine the central value and the offset into an individual value.

Separate intercepts can be written using $\mu + \alpha_i$ or using $\alpha_i^\star = \mu + \alpha_i$.

Separate slopes can be written using $\beta + \beta_i$ or using $\beta_i^\star = \beta + \beta_i$.

ANCOVA

The classic analysis of covariance compares the single line model to the parallel lines model.

It takes as base model the linear relationship of the response and the covariate and then asks whether the treatments shift the mean response up or down (do we need the α_is?).

This achieves variance reduction, because the linear relationship models out some of the variability that would be residual variability if we ignored the covariate and just did ANOVA.

Covariate Adjusted Means

In the parallel lines model, the covariate adjusted means are

$$\hat{\mu} + \hat{\alpha}_i + \hat{\beta}\bar{x}_{\bullet\bullet}$$

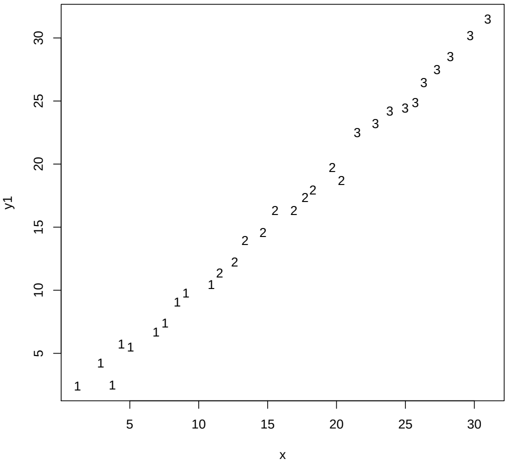
These are all evaluated at a common value of the covariate and differ according to the $\widehat{\alpha}_i$s.

The covariate adjusted means are almost always different than the treatment means because

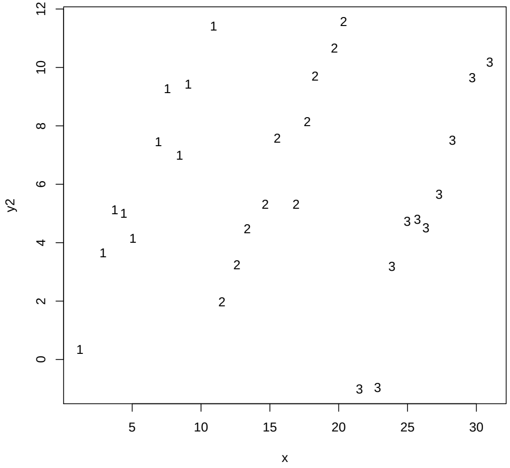
$$\overline{y}_{i\bullet} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}\bar{x}_{i\bullet}$$

Depending on the sign of $\hat{\beta}$ the pattern of $\bar{x}_{i\bullet}$, the covariate adjusted means can move up or down and be closer together or farther apart than the treatment means.

Most real world situations lie between the following two extremes.



Here the treatment means are different, but the covariate adjusted means will be very similar.



Here the treatment means are similar, but the covariate adjusted means will be very different.

Treatments affect covariates?

In ANCOVA situations, there is some variance that can only be explained by treatment differences, some variance that can only be explained by the covariate, and some variance that can be explained by either.

(The overlapping bit combines the slope and the covariate differences.)

The usual ANCOVA essentially assumes that any differences we see between treatments in covariate means are just random noise. This means that the overlapping variability should be attributed to the covariate, not the treatment.

But what if the treatments affect the covariates?

A classic example. Looking at the effect of some treatments on height growth of wheat plants. Experimental units are pots planted with a fairly large number of wheat seeds and given a treatment. Seeds sprout and grow, and then we measure height.

However, there is competition, so the more seeds sprout, the shorter the plants will be.

What if the treatment affects germination?

Adjusted covariates are

$$\tilde{x}_{ij} = x_{ij} - \overline{x}_{i\bullet}$$

These are the residuals from a model fitting the covariate as response to the treatments.

If we use adjusted covariates, then we get variance reduction, but we do no get covariance adjustment of the means. That is, the covariance adjusted means for this adjusted covariate are just the $\overline{y}_{i\bullet}$s.

Generalizations

More than one covariate.

Fancier designs for the treatments.