Topic 13: Covariance Analysis

- Covariable as a tool for increasing precision
- Carrying out a full ANCOVA
    - Testing ANOVA assumptions
- Happiness

Covariables as Tools for Increasing Precision

The eternal struggle of the noble researcher is to reduce the amount of variation that is unexplained by the experimental model (i.e. *error*). Blocking is one way of doing this, of reducing the variation-within-treatments (MSE) so that one can better detect differences among treatments (MST). Blocking is a powerful general technique in that it allows the researcher to reduce the MSE by allocating variation to block effects *without necessarily knowing the exact reason* for block-to-block differences. The price one pays for this general reduction in error is twofold: 1) Reduction in df_{error}, and 2) Constraint of the experimental design, as the blocks, if they are complete, must be crossed orthogonally with the treatments under investigation.

Another way of reducing unaccounted-for variation is through the use of covariables (or concomitant variables). Covariables are specific, observable characteristics of individual experimental units whose continuous values, while independent of the treatments under investigation, are strongly correlated to the values of the response variable in the experiment. Though it may sound a little confusing at first, the concept of covariables is quite intuitive. Consider the following example:

Baking potatoes takes forever, so a potato breeder decides to develop a fast-baking potato. He makes a bunch of crosses and carries out a test to compare baking times. While he hopes there are differences in baking times among the new varieties he has made, he knows that the *size* of the potato, regardless of its variety, will affect cooking time immensely. In analyzing baking times, then, the breeder must take into account this size factor. Ideally, he would like to compare the baking times *he would have measured had all the potatoes been the same size* – by using size as a covariable, he can do just that.

Unlike blocks, covariables are specific continuous variables that must be measured on each and every experimental unit, and their correlation to the response variable must be determined, in order to be used. One advantage over blocks is that covariables do not need to satisfy any orthogonality criteria relative to the treatments under consideration. But a valid ANCOVA (Analysis of Covariance) does require the covariable to meet several assumptions.

Impressing your friends by carrying out a full ANCOVA

Example 1 *ST&D p435 [Lab10ex1.R]*

This example illustrates the correct way to code for an ANCOVA, to test all ANCOVA assumptions, and to interpret the results of such an analysis.

In this experiment, eleven varieties of lima beans (Varieties 1 – 11) were evaluated for differences in ascorbic acid content (response Y, in mg ascorbic acid / 100 g dry weight). Recognizing that differences in maturity at harvest could affect ascorbic acid content *independent of variety*, the researchers measured the percent dry matter of each variety at harvest (X) and used it as a covariable (i.e. a proxy indicator of maturity).

So, in summary:

Treatment	Variety (1 – 11)
Response variable	Y (ascorbic acid content)
Covariable	X (% dry matter at harvest)

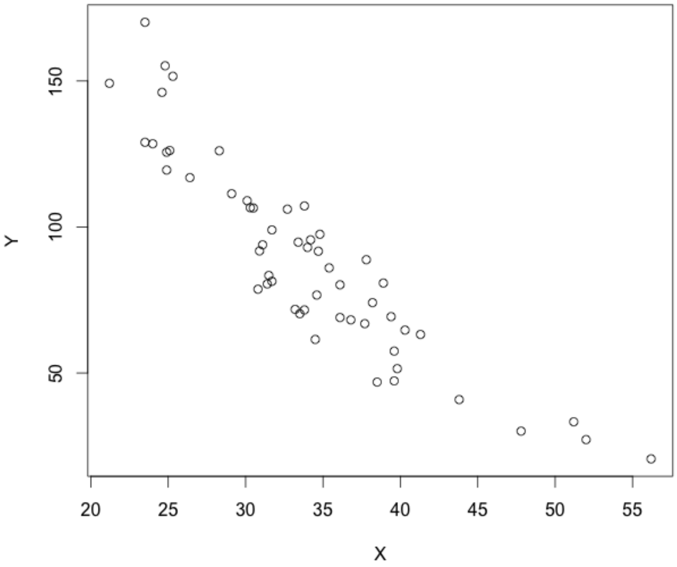
The experiment was organized as an RCBD with five blocks and one replication per block-treatment combination.

Block	Variety	X	Y
1	1	34	93
1	2	39.6	47.3
1	3	31.7	81.4
…	…	…	…
5	9	36.8	68.2
5	10	24.6	146.1
5	11	43.8	40.9

Coding, Results, and Explanations

1. The General Regression

```
plot(Y ~ X, lima_dat)
reg_mod<-lm(Y ~ X, lima_dat)
anova(reg_mod)
summary(reg_mod)
```



The above plot of Y vs. X indicates a clear negative correlation between the two variables, a correlation that is confirmed by the following regression analysis results:

```
Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1  51258   51258  254.44 < 2.2e-16 ***
Residuals 53  10677     201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 231.4084     9.1356   25.33   <2e-16 ***
X            -4.1925     0.2628  -15.95   <2e-16 ***
Multiple R-squared:  0.8276
```

There is a strong negative correlation (p < 2e-16, slope = -4.19) between ascorbic acid content and % dry matter at harvest. In fact, over 82% of the variation in ascorbic acid content can be attributed to differences in % dry matter. It is a result like this that suggests that an ANCOVA may be appropriate for this study.

2. The ANOVA with X (the covariable) as the response variable

```
anovaX_mod<-lm(X ~ Block + Variety, lima_dat)
anova(anovaX_mod)

Response: X
          Df  Sum Sq Mean Sq F value    Pr(>F)
Block      4  367.85  91.963  9.6384 1.486e-05 ***
Variety   10 2166.71 216.671 22.7087 1.851e-13 ***
Residuals 40  381.65   9.541
```

Depending on the situation, this ANOVA on X (the covariable) may be needed to test the assumption of independence of the covariable from the treatments. The two possible scenarios:

1. The covariable (X) is measured *before* the treatment is applied. In this case, the assumption of independence is satisfied *a priori*. If the treatment has yet to be applied, there is no way it can influence the observed values of X.

2. The covariable (X) is measured *after* the treatment is applied. In this case, it is possible that the treatment itself influenced the covariable (X); so one must test directly for independence. In this case, the null hypothesis is that there are no significant differences among treatments for the covariable. (i.e. Variety in this model should be NS).

Here we are in Situation #2. The "treatment" is the lima bean variety itself, so it was certainly "applied" before the % dry matter was measured. Here H_0 is soundly rejected ($p < 0.0001$), so *we should be very cautious* when drawing conclusions.

And how might one exercise caution exactly?

The danger here is that the detected effect of Variety on ascorbic acid content is really just the effect of differences in maturity. If the ANOVA is significant and the ANCOVA is not, you can conclude that difference in maturity is the driving force behind ascorbic acid differences. If, on the other hand, your ANOVA is not significant and your ANCOVA is, you can conclude that the covariable successfully eliminated an unwanted source of variability that was obscuring your significant differences among Varieties.

3. The ANOVA with Y as the response variable

```
anovaY_mod<-lm(Y ~ Block + Variety, lima_dat)
anova(anovaY_mod)
```

```
Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
Block      4   4969  1242.2  8.3549 5.363e-05 ***
Variety   10  51018  5101.8 34.3135 < 2.2e-16 ***
Residuals 40   5947   148.7
```

Multiple R-squared: 0.904

Looking at this ANOVA on Y (the response variable), there appears to be a significant effect of Variety on ascorbic acid content; but we will wait to draw our conclusions until we see the results of the ANCOVA (see comments in the yellow box on the previous page).

Incidentally, to calculate the relative precision of the ANCOVA to the ANOVA exactly, the **bold and red numbers** from the individual ANOVAs above will be used.

4. Testing for Homogeneity of Slopes

```
slopes_mod<-lm(Y ~ Block + Variety + X + Variety:X, lima_dat)
anova(slopes_mod)
```

```
Response: Y
            Df Sum Sq Mean Sq  F value    Pr(>F)
~~Block        4   4969  1242.2  27.3019 1.830e-09 ***~~
~~Variety     10  51018  5101.8 112.1279 < 2.2e-16 ***~~
~~X            1   3744  3743.8  82.2824 5.746e-10 ***~~
Variety:X   10    884    88.4   1.9428   0.07957 .
Residuals   29   1319    45.5
```

With this model, we are checking the homogeneity of slopes across treatments, assuming that the block effects are additive and do not affect the regression slopes. The interaction test is NS at the 5% level (p = 0.0796), so the assumption of slope homogeneity across treatments is not violated. Good. *This interaction must be NS in order for us to carry out the ANCOVA.*

5. The ANCOVA

```
#library(car)
ancova_mod<-lm(Y ~ Block + Variety + X, lima_dat)
Anova(ancova_mod, type = 2)
```

```
Anova Table (Type II tests)
Response: Y
          Sum Sq Df F value    Pr(>F)
Block      756.4  4  3.3469   0.01895 *
Variety   7457.6 10 13.1996 1.110e-09 ***
X         3743.8  1 66.2642 6.157e-10 ***
Residuals 2203.5 39
```

Multiple R-squared: 0.9644
MSE = 2203.5 / 39 = ***56.5***

Here we find significant differences in ascorbic acid content among varieties *besides those which correspond to differences in percent dry matter at harvest.* A couple of things to notice:

1. The model df has increased from 14 to 15 because X is now included as a regression variable. Correspondingly, the error df has dropped by 1 (from 40 to 39).

2. Adding the covariable to the model increased R^2 from 90.4% to 96.4%, illustrating the fact that inclusion of the covariable helps to explain more of the observed variation, increasing our precision.

3. Because we are now adjusting the Block and Variety effects by the covariable regression, no longer do the SS of the individual model effects add up to the Model SS:

$$\begin{aligned} 756.4 \text{ (Block SS)} + 7457.6 \text{ (Variety SS)} + 3743.8 \text{ (X SS)} &= 11957.8 \\ &\neq 59731.0 \text{ (Model SS)} \end{aligned}$$

When we allow the regression with harvest time to account for some of the observed variation in ascorbic acid content, we find that less variation is uniquely attributable to Blocks and Varieties than before. The covariable SS (3743.85) comes straight out of the ANOVA Error SS (5947.30 - 2203.45).

As with the ANOVAs before, the bold underlined term above (MSE_{ANCOVA}) is used in the calculation of relative precision, which can now be carried out:

$$\text{Effective MSE} = MSE_{Y-Adjusted} \cdot [1 + \frac{SST_X}{(t-1) \cdot SSE_X}]$$

$$= 56.5 \cdot [1 + \frac{2166.71}{10 \cdot 381.65}] = 88.576278$$

$$\text{Relative Precision} = \frac{MSE_{Y-Unadjusted}}{EffectiveMSE_{Y-Adjusted}} = \frac{148.7}{88.576278} = 1.68$$

In this particular case, we see that each replication with the covariable is as effective as 1.68 without; so the ANCOVA is 1.68 times more precise than the ANOVA ['Precise' in this sense refers to our ability to detect the *real* effect of Variety on ascorbic acid content (i.e. our ability to isolate the effects of Variety exclusive of the effects on dry matter content at harvest)].

Testing ANOVA Assumptions

To test the ANOVA assumptions for this model, we rely on an important concept: *An ANCOVA of the original data is equivalent to an ANOVA of regression-adjusted data.* In other words, if we define a regression-adjusted response variable Z:

$$Z_i = Y_i - b \cdot (X_i - \overline{X})$$

then we will find that an ANCOVA on Y is equivalent to an ANOVA on Z. Let's try it:

6. Find beta and Xmean; then create Z

```
summary(ancova_mod)
mean(lima_dat$X)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 193.0462    13.3420  14.469  < 2e-16 ***
....
X            -3.1320     0.3848  -8.140 6.16e-10 ***

> mean(lima_dat$X)
[1] 33.98727
```

```
lima_dat$Z<-lima_dat$Y + 3.1320*(lima_dat$X - 33.98727)
```

A Comparison of Means...

The table on the next page shows the adjusted Variety means generated by the lsmeans() function in the ANCOVA alongside the normal Variety means:

```
lima_lsm <- lsmeans(ancova_mod, "Variety")
adj_means <- aggregate(lima_dat$Z, list(lima_dat$Variety), mean)
```

Variety	ANCOVA (Y) LSMeans	ANOVA (Z) Means
1	92.587327	92.587327
2	79.116425	79.116425
3	78.103108	78.103108
4	84.530117	84.530117
5	95.983054	95.983054
6	97.506844	97.506844
7	99.978678	99.978678
8	72.044748	72.044748
9	81.146722	81.146722
10	122.783843	122.783843
11	74.319134	74.319134

Pretty impressive. This is striking confirmation that our imposed adjustment (Z) accounts for the covariable just as the ANCOVA does.

A Comparison of Analyses…

Compare the results of this ANCOVA (Y) table:

```
Anova Table (Type II tests)
Response: Y
          Sum Sq Df F value    Pr(>F)
Block      756.4  4  3.3469   0.01895 *
Variety   7457.6 10 13.1996 1.110e-09 ***
X         3743.8  1 66.2642 6.157e-10 ***
Residuals 2203.5 39
```

…and with those of this ANOVA (Z) table:

```
Response: Z
          Df  Sum Sq Mean Sq F value    Pr(>F)
Block      4   768.3  192.08  3.4869   0.01556 *
Variety   10 10984.6 1098.46 19.9406 1.528e-12 ***
Residuals 40  2203.5   55.09
```

Dependent Variable: Z

A couple of things to notice:

- **a.** Error SS is exactly the same in both (2203.5), indicating that the explanatory powers of the two analyses are the same.
- **b.** The F values have shifted, but this is understandable due in part to the shifts in df.
- **c.** The p-values are very similar in both analyses, giving us the same levels of significance as before.

The reason it is important to show the general equivalence of these two approaches is because, now that we have reduced the design to a simple RCBD, testing ANOVA assumptions is straightforward.

7. Normality of Residuals

```
anovaZ_mod<-lm(Z ~ Block + Variety, lima_dat)
lima_dat$anovaZ_resids <- residuals(anovaZ_mod)
shapiro.test(lima_dat$anovaZ_resids)

	Shapiro-Wilk normality test

data:  lima_dat$anovaZ_resids
W = 0.9709, p-value = 0.2016  NS
```

Our old friend Shapiro-Wilk is looking good.

8. Homogeneity of Variety Variances

```
#library(car)
leveneTest(Z ~ Variety, data = lima_dat)

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group 10   0.677 0.7395  NS
```

Lovely, that.

9. Tukey Test for Nonadditivity

```
lima_dat$anovaZ_preds <- predict(anovaZ_mod)
lima_dat$sq_anovaZ_preds <- lima_dat$anovaZ_preds^2
tukeyZ_mod<-lm(Z ~ Block + Variety + sq_anovaZ_preds, lima_dat)
anova(tukeyZ_mod)

Response: Z
                Df  Sum Sq Mean Sq F value    Pr(>F)
~~Block            4   768.3  192.08  3.4299   0.01701 *~~
~~Variety         10 10984.6 1098.46 19.6147 2.977e-12 ***~~
sq_anovaZ_preds  1    19.4   19.39  0.3462   0.55967  NS
Residuals       39  2184.1   56.00
```

And there we go. Easy.

The Take Home Message

To test the assumptions of a covariable model, adjust the original data by the regression and proceed with a regular ANOVA. In fact, adjusting the original data by the regression and proceeding with an ANOVA is a valid approach to the overall analysis. [*Note, however, that one first needs to demonstrate homogeneity of slopes to be able to use a common slope to adjust all observations.*]

10. Tukey Separation of Adjusted Means

As a final flourish, now that we see that the data meet all the assumptions of ANCOVA, we can dive into a means separation of the Varieties. To do this, remember it is necessary to use the lsmeans library contrast() function!

```
#library(lsmeans)
lima_lsm <- lsmeans(ancova_mod, "Variety")
contrast(lima_lsm, method = "pairwise", adjust = "tukey")
```

This little bit of code produces the following beast of a table that must then be combed through and organized into a standard ranked means separation table.

```
contrast    estimate        SE df t.ratio p.value
 1 - 2     13.470902  6.718773 39   2.005  0.6467
 1 - 3     14.484219  4.775527 39   3.033  0.1212
 1 - 4      8.057210  4.992092 39   1.614  0.8668
 ...             ...       ...  .     ...     ...
 9 - 10   -41.637122  5.872321 39  -7.090  <.0001
 9 - 11     6.827588  4.761519 39   1.434  0.9318
 10 - 11   48.464709  6.034373 39   8.031  <.0001

Results are averaged over the levels of: Block
P value adjustment: tukey method for a family of 11 means
```

Last lab -- congratulations!