13.5 Assumptions of ANCOVA

The assumptions of analysis of covariance are:

1. **The X's are fixed, measured without error, and independent of treatments.**

   This means that inferences are appropriate for interpolated (i.e. within range) rather extrapolated (outside range) values, that the measurement error for X is trivial relative to the observed variation in X, and that the treatments themselves do not affect the X values.

2. **The regression of Y on X, after accounting for treatment differences, is linear and independent of treatments.**

   This means that the regression is assumed to be approximately linear within the given range of X values and that the slopes of the regressions within each of the treatment levels are not significantly different from one another. Note that a linear relation is often a reasonably good approximation even in nonlinear cases, provided the values of the independent variable do not span too wide a range.

3. **The residuals are normally and independently distributed with zero mean and a common variance.**

   These are the normal assumptions of ANOVA.

13.5.1 Independence of X values from the treatments

If the covariable is measured **before** the treatments are applied, like in the previous oyster example, the independence of the treatments and the concomitant variable is satisfied by definition. However, if the concomitant variable is measured **after** the experiment, the independence of the covariable and the treatments should be examined.

An ANOVA which uses the covariable (X) as the response variable is appropriate to test this hypothesis. The null hypothesis is that there are no significant differences in X among treatment levels, and we expect to find no significant differences in order to be able to perform a standard covariance analysis.

CRD: `anova(lm(X ~ Trtmt, data_dat))`
RCBD: `anova(lm(X ~ Block + Trtmt, data_dat))`

The following code is included only as an example, because the test is not required in the oyster example.

From the oyster growth study:

```
#Testing for independence of X from Trtmt effects
oyster_X_mod<-lm(Initial ~ Trtmt, oyster_dat)
anova(oyster_X_mod)
```

The output:

```
Response: Initial
          Df  Sum Sq Mean Sq F value   Pr(>F)
Trtmt      4 176.79  44.198   4.985 0.009299 **
Residuals 15 133.00   8.866
```

In this case, we see that the differences in initial weight are highly significant. When the treatments are applied *after* measuring the covariable (i.e. when the covariable is not affected by the treatments) and the ANOVA of the covariable is found to be significant, the selected covariable will most likely have an effect on the final results. This is due to the fact that the experimental units across the treatments are not very uniform and adjusting for this nonuniformity will most likely increase the precision of tests.

ANCOVA can be used where the *X* values differ among treatments, but results should be interpreted with caution.

13.5.2 Test for heterogeneity of slopes

ANCOVA assumes homogeneity of covariate regression coefficients. This is ANCOVA's "equality of regressions" or "homogeneity of regressions" assumption. Because a single slope is used to adjust all observations in the experiment, the covariate coefficients (i.e. the slopes of the regression lines) must be the same (statistically) for each level of the categorical variable being analyzed (i.e. all slopes must estimate the same common slope β).

Said another way: The adjustment of the *Y* values using a single β for all treatments is based on the assumption that there is a constant regression relationship among groups. The test for heterogeneity of slopes tests the validity of this assumption; that is, it tests whether or not the regression coefficients are constant over groups.

The null hypothesis of this test is H_o: $\beta_1 = \beta_2 = ... = \beta_i$

A regression relationship that differs among treatment groups reflects an **interaction between**

the treatment groups and the independent variable or covariate. In fact, to test the assumption of homogeneity of slopes, one simply specifies and analyzes this phenomenon as an interaction. For example:

```
#Testing for homogeneity of slopes
oyster_slopes_mod<-lm(Final ~ Trtmt + Initial + Trtmt:Initial, oyster_dat)
anova(oyster_slopes_mod)
```

In this code, the expression "Trtmt:Initial" probes the treatment*regression interaction within the overall ANCOVA model. This factor in the model produces the appropriate statistics for estimating the differences among regressions across the different levels of Trt.

The output:

```
Response: Final
               Df  Sum Sq Mean Sq  F value    Pr(>F)
Trtmt           4 198.407  49.602 175.0233 3.284e-09 ***
Initial         1 156.040 156.040 550.5987 4.478e-10 ***
Trtmt:Initial   4   1.388   0.347   1.2247    0.3602  NS
Residuals      10   2.834   0.283
```

The row Trtmt:Initial presents the SS due to differences in regression coefficients among the groups specified by the classification variable Trtmt. If $p > 0.05$ for this interaction, we do not reject the hypothesis of homogeneity of slopes and are justified in proceeding with a standard analysis of covariance.

In more complex models, a similar strategy is followed. That is, to test for homogeneity of slopes across all levels of the factor(s) being tested, you need to run the standard ANCOVA model, to which the Trtmt:X interaction is added. For example, in an RCBD with one treatment factor (Trtmt) and a covariable (X):

```
#Testing for homogeneity of slopes (RCBD)
slopes_mod<-lm(Y ~ Block + Trtmt + X + Trtmt:X, data_dat)
anova(slopes_mod)
```

13.5.3 Normality of residuals

To test for normality of residuals, simply request the residuals from the ANCOVA model and perform a Shapiro-Wilk test on those residuals, as seen previously:

```
#Testing for normality of residuals (ANCOVA)
oyster_ancova_mod<-lm(Final ~ Trtmt + Initial, oyster_dat)
oyster_dat$ancova_resids <- residuals(oyster_ancova_mod)
shapiro.test(oyster_dat$ancova_resids)
```

Incidentally, this is equivalent to performing a Shapiro-Wilk test on the residuals of the ANOVA on the adjusted variable Z, as shown in the table below.

```
#Testing for normality of residuals (ANOVA on Z)
oyster_anovaZ_mod<-lm(Z ~ Trtmt, oyster_dat)
oyster_dat$anovaZ_resids <- residuals(oyster_anovaZ_mod)
shapiro.test(oyster_dat$anovaZ_resids)
```

	Trtmt	Rep	Initial	Final	Z	ancova_resids	anovaZ_resids
1	1	1	27.2	32.6	31.04022	0.887108538	0.887108538
2	1	2	32.0	36.6	29.84096	-0.312154592	-0.312154593
3	1	3	33.0	37.7	29.85778	-0.295334411	-0.295334412
4	1	4	26.8	31.0	29.87349	-0.279619535	-0.279619534
5	2	1	28.6	33.8	30.72377	0.606468758	0.606468758
6	2	2	26.8	31.7	30.57349	0.456192432	0.456192432
7	2	3	26.5	30.7	29.89845	-0.218853622	-0.218853622
8	2	4	26.8	30.4	29.27349	-0.843807568	-0.843807568
9	3	1	28.6	35.2	32.12377	0.071439716	0.071439715
10	3	2	22.4	29.1	32.73948	0.687154592	0.687154593
11	3	3	23.2	28.9	31.67294	-0.379389263	-0.379389262
12	3	4	24.4	30.2	31.67312	-0.379205045	-0.379205045
13	4	1	29.3	35.0	31.16554	-0.339141979	-0.339141980
14	4	2	21.8	27.0	31.28939	-0.215293338	-0.215293337
15	4	3	30.3	36.4	31.48236	-0.022321798	-0.022321799
16	4	4	24.3	30.5	32.08144	0.576757115	0.576757115
17	5	1	20.4	24.6	30.40584	0.008271928	0.008271928
18	5	2	19.6	23.4	30.07239	-0.325184217	-0.325184217
19	5	3	25.1	30.3	31.01490	0.617326779	0.617326778
20	5	4	18.1	21.8	30.09716	-0.300414489	-0.300414489

13.5.4 Homogeneity of variances

Recall that Levene's Test is only defined for one-way ANOVAs. To test for homogeneity of variances in an ANCOVA, therefore, it is customary to adjust the response variable (Y) according to its regression on X and perform a Levene's Test on the adjusted variable (Z):

```
# Find beta and mean X...
oyster_ancova_mod<-lm(Final ~ Trtmt + Initial, oyster_dat)
summary(oyster_ancova_mod)
mean(oyster_dat$Initial)

# Create Z
oyster_dat$Z<-oyster_dat$Final - 1.08318*(oyster_dat$Initial - 25.76)

# Test homogeneity of variances
#library(car)
leveneTest(Z ~ Trtmt, data = oyster_dat)
```

13.5.5 Additivity of main effects

Though unnecessary in the oyster example, note that in an RCBD with one replication per block-treatment combination, the adjusted Z values should also be used for the Tukey Test for Non-additivity. Example code:

```
#Testing for additivity of main effects
anovaZ_mod<-lm(Z ~ Block + Trtmt, data_dat)

data_dat$anovaZ_preds <- predict(anovaZ_mod)
oyster_dat$sq_anovaZ_preds <- oyster_dat$anovaZ_preds^2

tukeyZ_mod<-lm(Z ~ Block + Trtmt + sq_anovaZ_preds, data_dat)
anova(tukeyZ_mod)
```

13.6 Increase in precision due to covariance (ST&D 17.6)

To test the effectiveness of the covariate as a means of error control, a comparison can be made between the experimental errors, with and without the covariance adjustment. In the oyster example:

ANOVA of Y:	MSE = 10.68417	$df_{error} = 15$
ANCOVA of Y:	MSE = 0.30159	$df_{error} = 14$

To make the comparison between these two errors, the second value (ANCOVA MSE) must be adjusted upward to allow for sampling error in the regression coefficient. The appropriate adjustment involves the TRT SS (176.793) and the error SS (132.995) from the ANOVA of X. The **effective ANCOVA MSE**, adjusting for sampling error in X, is therefore given by:

$$MSE_{ANCOVA,Y}\left[1+\frac{SST_{ANOVA,X}}{(t-1)SSE_{ANOVA,X}}\right]=0.30159\left[1+\frac{176.793}{4*132.995}\right]=0.402$$

So, an estimate of the relative precision is

$$RE_{ANCOVA:ANOVA}=\frac{MSE_{ANOVA}}{\hat{MSE}_{ANCOVA}}=\frac{10.68417}{0.402}=26.6$$

This indicates that each replication, adjusted for the effect of the covariable, is as effective as 26.6 replications without such adjustment. The ANCOVA in this particular example is 26.6 times more precise than the ANOVA in detecting the effects of the treatments.

13.7 Comparison between ANCOVA and ANOVA of ratios

An experiment was conducted to study the effect of stress on the presence of enzyme A in the liver (Bulletin de la Societé de Chimie Biologique, 1954). The researcher measured the total activity of enzyme A (Variable "A") from liver homogenates of 10 control and 10 shocked animals. He also measured the total amount of nitrogen in the homogenates (Variable "N") as an indicator of total enzyme activity (i.e. total protein content) in the liver. He took this measure as a way to correct the measurements for A by total enzyme activity within each animal. Since he knew that A is correlated with N (but did not know about ANCOVA), he decided to analyze the ratio (A/N), the activity of enzyme per unit of protein, as his response variable.

The data:

Control animals			Shocked animals		
N	**A**	**A/N**	**N**	**A**	**A/N**
84	76	90.4	122	108	88.5
28	38	133.9	98	158	161.2
166	72	43.4	115	58	50.0
98	64	65.3	86	65	75.5
105	53	50.0	69	40	58.0
84	28	32.8	86	65	75.5
72	31	43.0	102	82	80.3
80	28	34.3	112	94	84.1
84	28	32.7	98	65	66.3
105	49	46.1	74	76	102.7

ANOVA of the variable (A/N) **$R^2 = 0.16$**

```
Response: A/N
           Df  Sum Sq Mean Sq F value  Pr(>F)
Group       1  3535.4  3535.4  3.5123 0.07724 NS
Residuals  18 18118.5  1006.6
```

The ANOVA indicates that there are no significant differences between treatments (p = 0.0772). This result is not unexpected when you reognize the large variance within groups for variable A/N. Note, for example, the big difference between the extreme values of the control group (32.7 and 133.9).

ANCOVA of the variable A, using N as a covariable **$R^2 = 0.99$**

```
Response: A
           Sum Sq Df F value  Pr(>F)
Group      5108.8  1  7.9807 0.01167 *
N          2162.6  1  3.3783 0.08360 .
Residuals 10882.4 17
```

In contrast to the result from the ANOVA of (A/N), the ANCOVA of A, using N as a covariable, shows significant differences between treatments (p = 0.0117). The increased precision of the ANCOVA is evident, but *why* is ANCOVA the superior analysis?

The use of ANOVA to analyze ratios $Z = Y/X$ is not correct. Both X and Y, being random variables, exhibit random variation; consequently, the derived variable Z will exhibit a variance that is some combination of the variances of X and Y. Variation in the numerator (Y) affects Z in a linear way, but variation in the denominator affects Z in a hyperbolic way ($Z = 1/X$ is the

equation of a hyperbole). This means that the magnitude of the error of Z depends not only on the error of *X* but also on the absolute value of *X*, with the error being higher for low values of *X*. Such a relationship affects the homogeneity of variances.

The correct way to analyze such ratios is an ANCOVA of the numerator, using the denominator as the covariable.

13.8 Uses of ANCOVA

The most important uses of the analysis of covariance are:

1. To control error and thereby to increase precision.
2. To adjust treatment means of the dependent variable for differences in the values of corresponding independent variables.
3. To assist in the interpretation of data, especially with regard to the nature or mechanism of treatment effects.
4. To estimate missing data.

13.7.1 Error control

Reduction of the experimental error (MSE) is accomplished through experimental design or through the use of one or more covariates; and both strategies may be used simultaneously. Covariance can be used as an effective means of error control when variation in the dependent variable Y is partly attributable to variation in some continuous, independent variable X.

ANCOVA increases the precision with which treatment effects can be measured by removing, via linear regression, certain recognized effects that cannot be or have not been controlled effectively by experimental design. For example, in a cattle-feeding experiment, to compare the effects of several rations on gain in weight, animals will always vary in initial weight. Now, if initial weight is correlated with gain in weight, a portion of the experimental error for gain can be the result of differences in initial weight. Through covariance analysis, this portion may be computed and eliminated from the experimental error, thereby increasing the researcher's ability to detect differences due to the rations.

13.7.2 Adjustment of treatment means

When observed variation in *Y* is partly attributable to variation in *X*, variation among treatment *Y* means will also be affected by variation among treatment *X* means. For valid comparisons, the treatment *Y* means should be adjusted to make them *the best estimates of what they would have been* had all treatment *X* means been the same.

For illustration, consider canning peas. This crop increases rapidly in yield as the peas progress in maturity. In a trial to evaluate the yields of different varieties, however, it is very difficult

(read: impossible) to harvest all the varieties at the same state of maturity. Given the strong dependence of yield on maturity, therefore, an analysis of yields unadjusted for maturity differences may have little value. Worse, such an analysis may lead to completely wrong conclusions. In contrast, a comparison of yields adjusted for maturity differences (i.e. using maturity as a covariable) could have great value.

In field experiments, yields are often adjusted for differences in plot productivity as determined by uniformity trials. A uniformity trial measures the yields from plots handled in a uniform manner prior to the execution of the main experiment. With annual crops, the increased precision resulting from the use of uniformity data rarely pays; however, with long-lived perennials such as tree crops, there is often much to be gained.

In animal feeding experiments, differences among treatment means unadjusted for the amount of food consumed may be due to differences in the nutritive value of the rations, to differences in the amounts consumed, or to both. If differences among mean gains in weight for the different rations are adjusted to a common food intake, the adjusted means will indicate whether or not the rations differ in nutritive value.

Here you can begin to see how covariables can help not only in increasing the ability to detect differences in treatments but also in understanding some of the reasons underlying *how* the treatments produce different effects.

13.7.3 Interpretation of data

As hinted above, covariance analysis often aids the experimenter in understanding the principles underlying the results of an investigation.

Recall that the adjusted treatment Y means estimate the values expected when the treatment X means are the same for all treatment groups. In cases where X is itself influenced by the treatments, this adjustment by X serves to remove part of the treatment effects on Y. In such situations, the results of the ANCOVA must be interpreted carefully.

As an example, consider a fertilizer trial on sugar beets, in which the researcher is interested in testing the effects of treatments on yield (Y). It is possible that the treatments may cause differences in stand (germination and establishment of the crop). In such cases, the analysis of yield (Y) adjusted for stand differences (X) will remove part of the treatment effect, thereby leading the experimenter to a wrong interpretation of the data. An ANCOVA can still supply useful information in this case, however. Total yield is a function of average weight per beet and of average stand per plot. Now, if stand is influenced by treatments, the ANCOVA of yield *adjusted for stand differences* would test specifically if the treatments affect *individual beet weights* rather than yield in general.

While an adjustment in observed yield that is proportional to the number of established plants is sometimes practiced in agricultural experiments, this procedure is not often recommended. This

is because yields are rarely proportional to the number of plants per plot, so such an approach usually results in an over-correction for the plots with smallest stands.

In situations where real differences among treatments for the independent variable do occur but are not the direct effect of the treatments, adjustment can be made without affecting interpretation. For example, consider a variety trial for which seed of the various varieties or strains was produced in different areas. Such seed may differ widely in germination rates, not because of inherent variety differences but as a result of the different environments in which the seed was produced. Consequently, differences in stand may occur even if planting rate is controlled. In this situation, the use of covariance for both error control and yield adjustment is warranted.

13.7.4. Estimation of missing data

Finally, the formulas given previously to estimate the values of missing data are based on the principle of minimizing the residual sum of squares. Such an approach, however, biases the treatment sum of squares upward. Covariance can be used to estimate missing values in a way that minimizes the residual sum of squares without biasing the treatment sum of squares. Such a procedure is simple to carry out though more difficult to describe than the previous procedure which required little more than a single formula.