

Topic 13. Analysis of Covariance (ANCOVA, ST&D Chapter 17)

13. 1. Introduction

The analysis of covariance is a technique that is occasionally useful for improving the precision of an experiment. Suppose that in an experiment with a response variable Y there is another variable, such as X , and that Y is linearly related to X . Furthermore, suppose that the experimenter cannot control X but can observe it along with Y . The variable X is called a **covariate** or **concomitant variable**. The analysis of covariance involves adjusting the observed response variable for the effect of the concomitant variable. If such an adjustment is not performed, the concomitant variable could inflate the error mean square and make true differences in the response due to treatments harder to detect. The concept is similar to the use of blocks to reduce the experimental error. However, when the blocking variable is a continuous variable, the delimitation of the blocks can be very subjective.

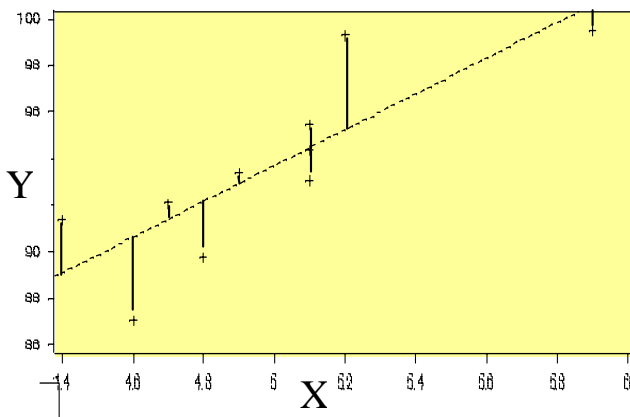
The ANCOVA uses information about Y that is contained in X in two ways:

- Variation in Y that is associated with X is removed from the error variance, resulting in more precise estimates and more powerful tests
- Group means of the Y variable are adjusted to correspond to a common value of X , thereby producing an equitable comparison of the groups.

Thus, the analysis of covariance is a method of adjusting for the effects of an uncontrollable nuisance variable. The procedure is a combination of analysis of variance and regression analysis. We will review briefly some concepts of regression analysis to facilitate the discussion of the analysis of covariance.

13. 2. Review of Regression concepts.

The equation of a straight line is $Y = a + bX$, where Y is the **dependent** variable and X is the **independent** variable. This straight line intercepts the Y axis at a so a is called the **intercept**. The coefficient b is the **slope** of the straight line: it represents the change in Y for each unit change in X . Any point (X, Y) on this line has an X coordinate, or **abscissa**, and a Y coordinate, or **ordinate**, whose values satisfy the equation.



Body weight, X	Food consumption, Y
4.6	87.1
5.1	93.1
4.8	89.8
4.4	91.4
5.9	99.5
4.7	92.1
5.1	95.5
5.2	99.3
4.9	93.4
5.1	94.4

$$Y_{estimated} = 55.26 + 7.69X$$

13. 2. 1. The principle of least squares: When a straight line is to be fitted to data consisting of (X, Y) pairs, one chooses the line that best fits the data. For each point we find its vertical distance from the straight line, square this distance, and then add together all the squared distances. Of all the lines that could possibly be drawn on the graph, the **best-fitting line** is the one that minimizes the sum of squared vertical deviations.

13. 2. 2. Residuals: The distance from a point to the straight line is a **residual** – the difference between the actual Y value and the Y value that the regression equation predicts. The residuals represent the behavior of Y that the independent variables don't account for—the error in the model.

13. 2. 3. Formulas to calculate a and b : Calculus provides the equations for the intercept a and the slope b that minimize the SS of the residuals:

$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{S(XY)}{SS(X)} \text{ and } a = \bar{Y} - b\bar{X}$$

13. 2. 4. Covariance: $S(XY)$ is called the **corrected sum of cross products**. Dividing $S(XY)$ by $n-1$ produces a statistic called the sample **covariance** between X and Y . If high values of Y are associated with high of X the covariance will be positive. If high values of Y are associated with low values of X the covariance will be negative. If there is no association the covariance will be close to zero.

13. 2. 5. Using SAS for regression analysis: PROC REG and PROC GLM can be used for regression analysis.

```
data STp254_r;
input x y @@;
cards;
4.6 87.1    5.1 93.1    4.8 89.8    4.4 91.4    5.9 99.5
4.7 92.1    5.1 95.5    5.2 99.3    4.9 93.4    5.1 94.4
;
proc glm; (Note that the independent variable is not included in a CLASS statement)
model y= x;
run;
```

Output:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	90.835510	90.835510	16.23	0.0038
Error	8	44.768490	5.596061		
Corrected Total	9	135.604000			

F: tests if the model as a whole accounts for a significant proportion of Y .

R-Square	C.V.	Root MSE	Y Mean
0.669859	2.528430	2.3656	93.560

R-Square: measures how much variation in Y the model can account for.

Parameter	Estimate	T for H0: Parameter=0	Pr > T	Std Error of Estimate
INTERCEPT	55.26328125	5.80	0.0004	9.53489040
X	7.69010417	4.03	0.0038	1.90873492

Estimates: Calculates the INTERCEPT ($a= 55.26$) and the slope ($b= 7.69$) and test if they are significantly different from 0.

13. 2. 6. ANOVA of the adjusted Y's

The MS_{error} (44.77) from the previous analysis represents the variation in Y (food consumption) that would have been obtained if all the animals used in the experiment had had the same initial (X) body weight.

In the following table each Y value is adjusted using the regression equation to a common X . Any value of X can be used to adjust the Y 's but the mean of the X (4.98) values is used as a representative value:

X	Y	<i>Adjusted Y</i> = $Y - b(X - \bar{X} \dots)$
4.6	87.1	90.02224
5.1	93.1	92.17719
4.8	89.8	91.18422
4.4	91.4	95.86026
5.9	99.5	92.4251
4.7	92.1	94.25323
5.1	95.5	94.57719
5.2	99.3	97.60818
4.9	93.4	94.01521
5.1	94.4	93.47719
$X_{\text{mean}} =$	4.98	
SSY	135.604	44.76849

The first adjusted value, 90.02224, is the food consumption expected for this animal if its initial body weight would have been 4.98. Note that the SS of the Y 's is similar to the Total SS of the previous ANOVA and that the SS of the adjusted Y 's is similar to the SS_{error} . The SS_{error} is the variation in food consumption that we would have found if all the animals used in the experiment had had the same weight (assuming that "b" was estimated without error).

Note the large reduction in the variation of the Y 's that is obtained when the variation due to the regression is eliminated.

13. 3. ANCOVA example

The analysis of covariance is illustrated below by data on the growth of oysters. The goal of this experiment is to determine

- if exposure to water heated artificially affects growth
- if the position in the water column (surface or bottom) affects growth

Four bags with ten oysters in each bag are randomly placed at each of 5 locations in the cooling water canal of a power-generating plant. Each location is considered a treatment: TRT1: cool-bottom, TRT2: cool-surface, TRT3: hot-bottom, TRT4: hot-surface, TRT5: control mid-depth and mid-temperature.

Each bag is considered to be one experimental unit. The oysters are cleaned and weighted at the beginning of the experiment and then again about one month later. The initial weight and the final weight are recorded for each bag.

The data (from SAS System for linear models) is included in the following SAS program.

```

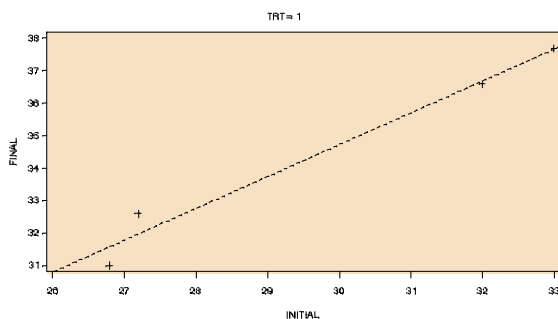
data oyster;
input trt rep initial final;
cards;
1 1 27.2 32.6
1 2 32.0 36.6
1 3 33.0 37.7
1 4 26.8 31.0
2 1 28.6 33.8
2 2 26.8 31.7
2 3 26.5 30.7
2 4 26.8 30.4
3 1 28.6 35.2
3 2 22.4 29.1
3 3 23.2 28.9
3 4 24.4 30.2
4 1 29.3 35.0
4 2 21.8 27.0
4 3 30.3 36.4
4 4 24.3 30.5
5 1 20.4 24.6
5 2 19.6 23.4
5 3 25.1 30.3
5 4 18.1 21.8
;
proc GLM;
Title 'One-way Anova';
class trt;
model final= trt;

proc GLM;
Title 'ANCOVA';
class trt;
model final= trt initial;
run; quit;

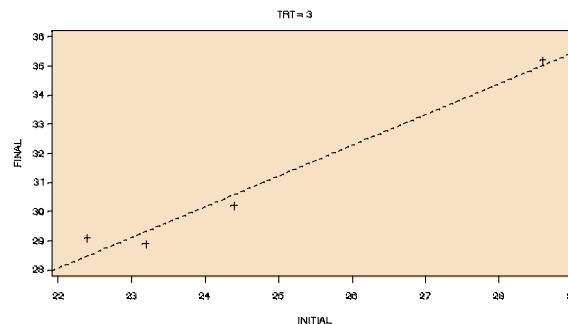
```

The CLASS statement specifies that TRT is a classification variable. The variable “**initial**” is the **covariate**. Note that “initial” is a continuous variable and is NOT

included in the CLASS statement. This is similar to the regression example in 13. 2. 4. A simple regression of Final versus Initial showed a large significant regression ($p < 0.01$, $r^2 = 0.95$). Note that the slopes of the regressions within the different treatments are similar. Examples of final vs. initial regressions for treatments 1 and 3 are shown below.



$$\text{TRT1: } Y = 5.24 + 0.98X$$



$$\text{TRT3: } Y = 4.82 + 1.06X$$

SAS Output:

One-way ANOVA

Dependent Variable: FINAL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
TRT (= Model)	4	198.40700	49.60175	4.64	0.0122
Error	15	160.26250	10.68417		
Corrected Total	19	358.66950			

R-Square	C.V.	Root MSE	FINAL Mean
0.553175	10.59706	3.2687	30.845

The simple one-way ANOVA discovers treatment differences ($p = 0.0122$) in final weight even when the initial weights are not considered

ANCOVA

Dependent Variable: FINAL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	354.44718	70.88944	235.05	0.0001
Error	14	4.22232	0.30159		
Corrected Total	19	358.66950			

R-Square	C.V.	Root MSE	FINAL Mean
0.988228	1.780438	0.5492	30.845

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TRT	4	198.40700	49.60175	164.47	0.0001
INITIAL	1	156.04018	156.04018	517.38	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRT	4	12.08936	3.02234	10.02	0.0005
INITIAL	1	156.04018	156.04018	517.38	0.0001

The Type I SS for TRT is the unadjusted treatment SS and is the same as the one found in the one-way ANOVA. If we subtract this SS from the Total SS we obtain the error SS for the simple one-way ANOVA ($358.6695 - 198.407 = 160.2625$).

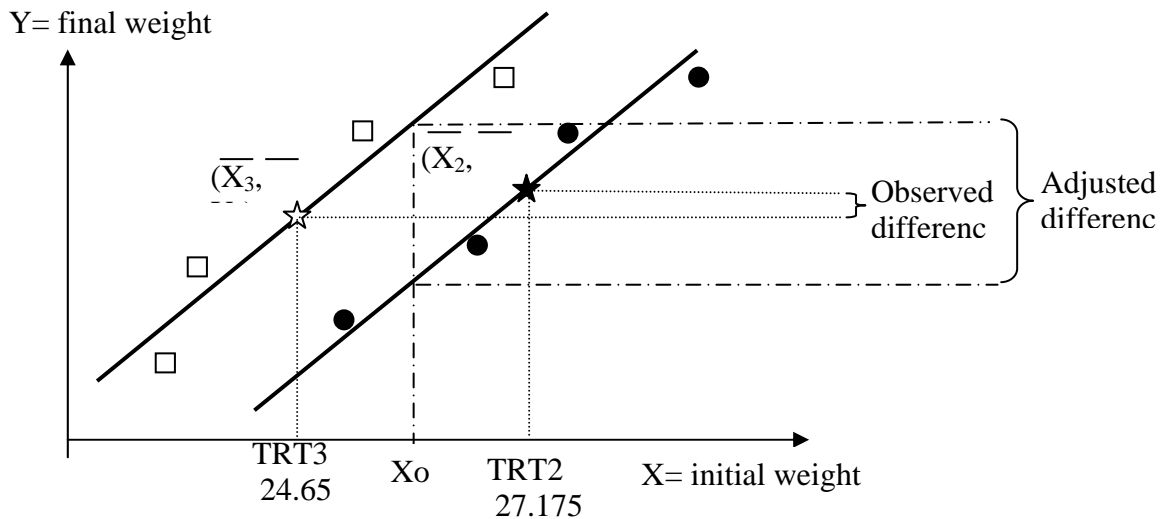
The Type III SS (12.089) is the **adjusted** treatment SS and enable us to test the treatment effects adjusted for all other factors included in the model.

Type III SS will produce the appropriate results for the ANCOVA.

Though this MS is smaller than the unadjusted TRT MS (49.60175) the reduction in the error term is even larger (from 10.684 to 0.30159). This allows an increase in the F statistic of the test from 4.642 in the simple one-way ANOVA to 10.02 in the ANCOVA. The power of the test for treatment differences increases when the covariate is included because most of the error in the simple ANOVA is due to variation in INITIAL values.

Finally, the INITIAL SS is used to test the significance of the regression between INITIAL and FINAL once the SS of INITIAL was adjusted for TRT.

13. 3. 1. Graphic interpretation of the ANCOVA example



The black circles represent the data from the four bags of treatment 2. The white squares represent the data from treatment 3. The mean final weight of treatment 3 (30.85) is slightly lower than the mean of treatment 2 (31.65).

For each treatment, variation in X is seen to contribute to variation in Y. Therefore, the distance between the initial weight averages of the oysters assigned to each treatment can contribute greatly to the difference between the final average weights. Treatment 3 started with an average weight of 24.65 and treatment 2 with an average weight of 27.175. If the treatments means had been observed from some common

average X , say X_o , then they would be comparable. Thus the need for adjusting treatment means is apparent.

This X_o can be thought as a common mean obtained after moving the values of treatment 3 upwards along the regression line and the values of treatment 2 downward along the regression line. The adjusted mean of treatment 3 is significantly larger than that of treatment 2.

13.3.2. Least squares means or adjusted means

If a MEANS statement is included in the previous example it will show the unadjusted treatment means of all continuous (non-CLASS) variables in the model. As discussed in the graphic example above, these means or their comparison are not strictly appropriate.

To be comparable, treatment means should be adjusted to make them the best estimates of what they would have been if all treatment independent means had been the same. These adjusted means can be calculated in SAS by using the LSMEANS (least-squares means) statement. This statement is the same one used to obtain adjusted means for the unbalanced two-way classification. These adjusted means are obtained using the following formula:

$$LS \bar{Y}_i = \bar{Y}_i - \hat{\beta}(\bar{X}_i - \bar{X}_{..})$$

The statement:

```
lsmeans trt /stderr pdiff adjust=tukey;
```

will print the estimated least-squares means followed by their standard errors (stderr option) and significance probabilities for all pairwise tests of treatment differences (pdiff option).

Note the large changes in going from the unadjusted to adjusted treatments means for the variable FINAL weight in the table below:

TRT	Unadjusted Means	Adjusted LS Means	Calculation
1	34.475	30.153	34.475 - 1.08318 (29.75 - 25.76)
2	31.650	30.117	31.650 - 1.08318 (27.18 - 25.76)
3	30.850	32.052	30.850 - 1.08318 (24.65 - 25.76)
4	32.225	31.504	32.225 - 1.08318 (26.43 - 25.76)
5	25.025	30.398	25.025 - 1.08318 (20.80 - 25.76)

These changes are due to the large treatment differences for the variable INITIAL. Some treatments, particularly TRT 5, received smaller oysters than other treatments. The coefficient $\beta=1.08318$ is a weighted average of the regression coefficients of FINAL on INITIAL, estimated separately for each of the five treatment groups. To obtain this coefficient add the SOLUTION option to the MODEL statement:

model final = trt initial/ **solution**;

The slope b obtained by the **solution** statement, is identical to the slope obtained by performing an ANOVA on both X and Y, calculating the residuals, and then running a regression of the **Y residuals on the X residuals**. This slope b can be used to create a new variable of adjusted values:

$$Z = Y - b * (X - \bar{X}_{..})$$

This ANOVA of adjusted values can be used to perform Levene's test of homogeneity of variances and Tukey tests of non additivity if the design is an RCBD with only one observation per treatment block combination.

13. 3. 3. Contrasts

The adjusted treatment means from the analysis of covariance can be analyzed further with four orthogonal contrasts implemented by the following CONTRAST statements:

```
proc GLM;
  class trt;
  model final= trt initial;
  contrast 'control vs. Treatment'    TRT -1 -1 -1 -1 4;
  contrast 'bottom vs. top'          TRT -1 1 -1 1 0;
  contrast 'cool vs, hot'            TRT -1 -1 1 1 0;
  contrast 'interactions depth*temp' TRT 1 -1 -1 1 0;
```

Output:

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
control vs. Treatmen	1	0.5200041	0.5200041	1.72	0.2103
bottom vs. top	1	0.3387907	0.3387907	1.12	0.3071
cool vs. hot	1	8.5910808	8.5910808	28.49	0.0001
interactions depth*T	1	0.2293415	0.2293415	0.76	0.3979
		9.6792171			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRT	4	12.08936	3.02234	10.02	0.0005

The output shows that the only significant difference is cool vs. hot. Although constructed to be orthogonal, these contrasts are not orthogonal to the covariable; hence, their sums of squares do not add to the adjusted treatment SS.

If the covariable is **not** included in the model, the same contrast statements produce completely different results!

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
control vs. Treatmen	1	169.36200	169.36200	15.85	0.0012
bottom vs. top	1	2.10250	2.10250	0.20	0.6637
cool vs. hot	1	9.30250	9.30250	0.87	0.3655
interactions depth*T	1	17.64000	17.64000	1.65	0.2183

The significance of the control vs. treatment contrast is due to the lower initial weight of the oysters placed in the control bags. In this case the contrast SS will add to the TRT SS (198.407). Note that the contrast statement will partition the TRT SS resulting for the specified model.

13. 4 ANCOVA model

The ANOVA model for a CRD is

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

The regression model is:

$$Y_i = \mu + \beta(X_i - \bar{X}_{..}) + \varepsilon_i$$

$$Y_{ij} = \mu + \tau_i + \beta(X_{ij} - \bar{X}_{..}) + \varepsilon_{ij}$$

The ANCOVA model for a CRD is:

The linear additive model for any given design is that for the ANOVA plus an additional term for the concomitant or independent variable. The variable being analyzed, Y , generally denotes the dependent variable, whereas the variable used in the control of error and adjustment of means, is denoted by X .

The rearranged formula

$$Y_{ij} - \beta(X_{ij} - \bar{X}_{..}) = \mu + \tau_i + \varepsilon_{ij}$$

indicates that the ANCOVA is a regular ANOVA of values that have been adjusted for regression on an independent variable.

13. 5 Assumptions of the ANCOVA

The assumptions necessary for the valid use of covariance are:

- a. The X 's are fixed, measured without error, and independent of treatments.

This means that inferences will be for the interpolated rather than extrapolated values, that the measurement error is trivial relative to the observed variation, and that the treatments itself will not affect the X values.

- b. The regression of Y on X after removal of the treatment differences is linear and independent of treatments.

This means that the regression is assumed to be approximately linear within the range of X values, and that the slopes of the regressions within the treatments are not significantly different. A linear relation is often a reasonably good approximation for a nonlinear relation provided the values of the independent variables do not cover too wide a range.

- c. The residuals are normally and independently distributed with zero mean and a common variance.

These are the normal assumptions for the validity of the F tests.

13. 5. 1. Independence of X values from the treatments

If the covariable is measured **before** the experiment, like in the previous oyster example (13.3) the independence of the treatments and the concomitant variable is always satisfied. However, if the concomitant variable is measured **after** the experiment the independence of the covariable and the treatments should be tested.

An analysis of variance of the covariable using the treatments as CLASS variable is appropriate to test this hypothesis. The null hypothesis is that there are no significant differences among treatments for the covariable. We expect to find no significant differences in order to be able to perform a standard covariance analysis.

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

```
proc glm;
  title 'Test for independence of treatments and covariable';
  class trt;
  model initial= trt;
run;
```

Dependent Variable: INITIAL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
TRT= Model	4	176.79300	44.19825	4.98	0.0093
Error	15	132.99500	8.86633		

R-Square	Coeff Var	Root MSE	initial Mean
0.570690	11.55916	2.977639	25.76

In this case the differences in initial weight are highly significant. When the application of the treatments after measuring the covariable ensures that the covariable is not affected by the treatments, and the ANOVA of the covariable is significant, the selected covariable will most likely have an effect in the final results.

Covariance can be used where the X values are affected but it must be interpreted with caution.

13. 5. 2. Test for heterogeneity of slopes.

Homogeneity of covariate regression coefficients. This is ANCOVA's "equality of regressions" or "homogeneity of regressions" assumption. The covariate coefficients (the slopes of the regression lines) are the same for each group formed by the categorical variables and measured on the dependent.

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 (assumption "b" 13. 5.). 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 is $H_0: \beta_1 = \beta_2 = \dots = \beta_i$

Regression relationships that differ among treatment groups actually reflect an **interaction between the treatment groups and the independent variables** or covariates. In fact the GLM procedure specifies and analyzes this phenomenon as an interaction. Thus, if you use the following statements, the expression $X*A$ produces the appropriate statistics for estimating different regressions of Y on X for the different classes specified by A .

For the previous example (Topic 13.3) the appropriate statements are

```
proc glm;
  title 'Test for heterogeneity of slopes';
  class trt;
  model final= trt initial trt*initial;
run; quit;
```

Output:

Dependent Variable: FINAL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	355.83549	39.53728	139.51	0.0001
Error	10	2.83401	0.28340		
Corrected Total	19	358.66950			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TRT	4	198.40700	49.60175	175.02	0.0001
INITIAL	1	156.04018	156.04018	550.60	0.0001
INITIAL*TRT	4	1.38831	0.34708	1.22	0.3602

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRT	4	1.696187	0.424047	1.50	0.2752
INITIAL	1	68.528923	68.528923	241.81	0.0001
INITIAL*TRT	4	1.388314	0.347079	1.22	0.3602

The last row INITIAL*TRT is an additional SS due to different regression coefficients for the groups specified by TRT. If this $p > 0.05$ we do not reject the hypothesis of homogeneity of slopes. In this case we accept the homogeneity of slopes and a standard covariance analysis is indicated.

In more complex models enter **all main effects of the factors and the covariate** and the interaction of the covariate(s) with the factor(s). These interaction effects should be non-significant if the homogeneity of regressions assumption is met.

For example in an RCBD with covariable X (ST&D page 449)

```
proc glm;
  title 'Test for heterogeneity of slopes';
  class block trt;
  model final= block trt X trt*X;
```

13. 5. 3. Analysis of residuals.

Similar statements to those used in previous analysis can be included to check the normality and homogeneity of variance of the residuals:

```
proc GLM;
  class trt;
  model final= trt initial;
  output out=check p=predi r=resi;

proc univariate data=check normal;
  var resi;

proc plot;
  plot resi*predi=trt;

run; quit;
```

13. 6. Increase in precision due to covariance (ST&D 17.6)

To test the effectiveness of covariance as a means of error control, a comparison is made of the variance of the treatment mean with and without the covariance adjustment.

For the previous examples the ANOVA on the unadjusted Y's has error MS 10.68417 with 15 df, and the ANCOVA on the adjusted Y's has error MS 0.30159 with 14 df (Topic 13.3). This last value must be adjusted upward to allow for sampling error in the regression coefficient. The adjustment involves the TRT SS (176.793) and the error SS (132.995) from an ANOVA on X (Topic 13. 15. 1.).

The **effective error MS** after adjustment for X is given by

$$MS_{\text{Error Adjusted Y}} \left[1 + \frac{\text{TRT } SS_{X \text{ variable}}}{(t-1) \text{ Error } SS_{X \text{ variable}}} \right] = 0.30159 \left[1 + \frac{176.793}{4 * 132.995} \right] = 0.402$$

An estimate of the relative precision is

$$MS_{\text{Error Unadjusted Y}} / \text{Effective } MS_{\text{Error Adjusted Y}} = 10.68417 / 0.402 = 26.6,$$

This indicates that each replications with the covariance is as effective as 26.6 without. The ANCOVA in this particular example is 26.6 times more precise than the ANOVA.

13. 7. Comparison between ANCOVA and ANOVA of ratios

A researcher wants to study the effect of stress on the presence of enzyme A in the liver (Bulletin de la Soci t  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 N (Variable "N") to correct the measurements of total enzyme activity using the total protein present in the liver. Since he knew that A is correlated with N (and he did not studied ANCOVA), he decided to analyze the ratio A/N or the activity of enzyme per unit of protein.

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 for the variable A/N.

Dependent Variable: A/N

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model = TRT	1	3650.4020	3650.4020	3.66	0.0719
Error	18	17970.7180	998.3732		
Corrected Total	19	21621.1200			

The ANOVA indicates that there are no significant differences between treatments. This result expected based on the large variance within groups for variable A/N. Note the large difference between the extreme values (control: 32.7 and 133.9)

ANCOVA for the variable A using N as covariable.

Dependent Variable: A

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8079.3923	4039.6961	6.31	0.0089
Error	17	10882.4077	640.1416		
Corrected Total	19	18961.8000			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRT	1	5108.8082	5108.8082	7.98	0.0117
N	1	2162.5923	2162.5923	3.38	0.0836

The ANCOVA shows significant differences between treatments. The increased precision of the ANCOVA is evident

The use of ANOVA to analyze ratios $Z = Y/X$ is not correct. Both X and Y have random variation, and consequently variation in Z will be the result of the combination of the variations of X and Y . Variation in the numerator (Y) affect Z in a lineal way but variation in the denominator affect Z in a hyperbolic way ($Z=1/X$ is the equation of a hyperbole). Variation below the X mean has a larger effect on Z than variation above the X mean. Moreover, the magnitude of the error of Z depends not only on the error of X but

also on the absolute value of X . The error is higher for low values of X . This clearly affects the homogeneity of variances.

The correct way to analyze these ratios is an ANCOVA for the numerator using the denominator of the ratio as covariable.

13. 8. Uses of ANCOVA (ST&D p 429)

The most important uses of covariance analysis are

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

13. 7. 1. Error control. Control of the MS error is accomplished by experimental design or by means of one or more covariates. Both methods may be used simultaneously. Covariance can be used as a method of reducing error when variation in the dependent variable Y is partly attributable to variation in the independent variable X .

The use of covariance to control error is a means of increasing the precision with which treatment effects can be measured by removing, by 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 assigned to any one block will 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. By covariance analysis, this portion, may be computed and eliminated from experimental error for gain.

13. 7. 2. To adjust treatment means When observed variation in Y is partly attributable to variation in X , variation in treatment Y means must also be affected by variation among treatment X means. To be comparable, treatment Y means should be adjusted to make them the best estimates of what they would have been if all treatment X means had been the same.

For illustration, consider canning peas. This crop increases rapidly in yield with increase in maturity. In a trial to evaluate yields of different varieties, it is difficult to harvest all at the same state of maturity. An analysis of yields unadjusted for differences in maturity may have little value. However, maturity can be used as a covariate. A comparison of yields adjusted for maturity differences would be more meaningful than a comparison among unadjusted yields.

In field experiments, yields can be adjusted for differences in plot productivity as determined by uniformity trials. A uniformity trial measures yields from plots handled in a uniform manner prior to the performance 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 unadjusted treatment means 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 indicate whether or not the rations differ in nutritive value.

Here covariance is getting at the principles underlying the results of the investigation by supplying information on the way in which the treatments produce effects.

13. 7. 3. Interpretation of data. Covariance analysis often aids the experimenter in understanding the principles underlying the results of an investigation.

If the independent variable is influenced by the treatments, the interpretation of the data is changed. This is so because the adjusted treatment means estimate the values expected when the treatment means for the independent variable are the same. Adjustment removes part of the treatment effects when means of the independent variable are affected by treatments.

In a fertilizer trial on sugar beets, the treatments may cause differences in stand. When stand, the independent variable, is influenced by treatments, the analysis of yield adjusted for stand differences removes part of the treatment effect and the experimenter may be misled in the interpretation of the data. An analysis of covariance can still supply useful information. Total yield is a function of average weight per beet and of stand. Now if stand is influenced by treatments, the analysis of covariance of yield adjusted for stand differences would indicate that treatments affect individual beet weights on the average.

An adjustment in proportion to the number of plants is sometimes practiced. This procedure is not recommended, because it usually results in an over-correction for the plots with smallest stand since yields are rarely proportional to the number of plants per plot. The analysis of covariance provides a more satisfactory and appropriate method of adjusting the experimental data.

In situations where real differences among treatments for the independent variable do occur but are not the direct effect of the treatments, adjustment is warranted. For example, consider a variety trial for which seed of the various varieties or strains has been produced in different areas. Such seed may differ widely in germination, not because of inherent differences but as a result of the environment in which it was grown. 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 Formulas given previously estimating missing data result in a minimum residual sum of squares. However, the treatment sum of squares is biased upward. The use of covariance to estimate missing plots results in a minimum residual sum of squares and an unbiased treatment sum of squares. The covariance procedure is simple to carry out though more difficult to describe than previous procedures which required little more than a formula.

Example of a complete SAS analysis of ANCOVA

```

data oyster;
input trt rep initial final;
Z= final-1.083179819*(initial-25.76);
cards;
1 1 27.2 32.6
1 2 32.0 36.6
1 3 33.0 37.7
1 4 26.8 31.0
2 1 28.6 33.8
2 2 26.8 31.7
2 3 26.5 30.7
2 4 26.8 30.4
3 1 28.6 35.2
3 2 22.4 29.1
3 3 23.2 28.9
3 4 24.4 30.2
4 1 29.3 35.0
4 2 21.8 27.0
4 3 30.3 36.4
4 4 24.3 30.5
5 1 20.4 24.6
5 2 19.6 23.4
5 3 25.1 30.3
5 4 18.1 21.8
;

proc GLM;
  title 'One-Way ANOVA for final and initial';
  class trt;
  model initial final= trt;

proc GLM;
  title 'general regression';
  model final= initial;

proc GLM;
  title 'ANCOVA';
  class trt;
  model final= trt initial / solution;
  output out=check p=predi r=resi;
  contrast 'control vs. Treatment' TRT -1 -1 -1 -1 4;
  contrast 'bottom vs. top' TRT -1 1 -1 1 0;
  contrast 'low vs, high' TRT -1 -1 1 1 0;
  contrast 'interactions depth*temp' TRT 1 -1 -1 1 0;
  lsmeans trt/ stderr pdiff adjust=tukey;

proc univariate data=check normal;
  title 'Normality of residuals';
  var resi;

proc plot;
plot resi*predi=trt;

proc glm;
  title 'Heterogeneity of slopes';
  class trt;
  model final= initial trt trt*initial;

proc GLM;
  title 'ANOVA adjusted Z';
  class trt;
  model Z= trt;
  output out=ZZ p=predZ r=resZ;
  means trt / hovtest= Levene;

run; quit;

```

The values for the adjusted Z are obtained from

The X mean is obtained from the one-way X-ANOVA (page 13.10)

The slope is obtained from the last value of the **solution** output.

The adjusted Z can be used to test homogeneity of variances of adjusted values

In an RCBD with one rep, you can perform Tukey test of non-additivity by

```
Proc GLM Data = ZZ;  
  Class Block trt  
  Model Z = Block trt predZ*predZ; *Tukey test non-additivity
```