Topic 13. Analysis of Covariance (ANCOVA) - Part I [ST&D Ch. 17]

13.1 Introduction

The analysis of covariance (ANCOVA) is a technique that can be useful for improving the precision of an experiment. Suppose that in an experiment with a response variable Y, there is another variable (X) such that Y is linearly related to (i.e. covaries with) X. Furthermore, suppose that the researcher cannot control X but can observe it along with Y. Such a variable X is called a **covariate** or a **concomitant variable**. The basic idea underlying ANCOVA is that precision in detecting the effects of treatments on Y can be increased by adjusting the observed values of Y for the effect of the concomitant variable. If such adjustments are not performed, the concomitant variable X could inflate the error mean square and make true differences in the response due to treatments harder to detect. The concept is very similar to the use of blocks to reduce the experimental error. But whereas the delimitation of blocks can be very subjective/arbitrary when the blocking variable is a continuous variable, controlling error in such cases via a covariable is straightforward.

The ANCOVA uses information about X in two distinct ways:

1. Variation in Y that is associated with variation in X is removed from the error variance (MSE), resulting in more precise estimates and more powerful tests.
2. Individual observations of Y are adjusted to correspond to a common value of X, thereby producing group means that are not biased by X, as well as equitable group comparisons.

A hybrid of ANOVA and linear regression analysis, ANCOVA is a method of adjusting for the effects of an uncontrollable nuisance variable. We will review briefly some concepts of regression analysis to facilitate this discussion.

13.2 Review of regression concepts

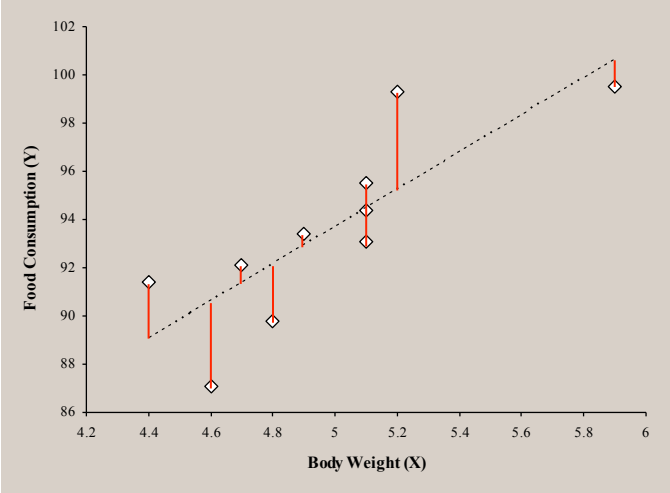
The equation of a straight line is $Y = a + bX$, where Y is the **dependent** variable and X is the **independent** variable. This straight line intercepts the Y axis at the value a, so a is called the **intercept**. The coefficient b is the **slope** of the straight line and represents the change in Y for each unit change in X (i.e. rise/run). Any point (X,Y) on this line has an X coordinate, or **abscissa**, and a Y coordinate, or **ordinate**, the pair of which satisfies the equation.

13.2.1 The principle of least squares

To find the equation of the straight line that best fits a dataset consisting of (X,Y) pairs, we use a strategy which relies on the concept of **least squares**. For each point in the dataset, we find its vertical distance from the putative best-fit straight line, square this distance, and then add together all the squared distances (i.e. vertical deviations). Of all the lines that could possibly be drawn through the scatter of data, the **line of best fit** is the one that minimizes this sum.

Example: Below is a scatterplot relating the body weight (X) of 10 animals to their individual food consumption (Y). The data are shown to the left.

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



13.2.2 Residuals

The vertical distance from an individual observation to the best-fit line is called the **residual** for that particular observation. These residuals, indicated by the solid red lines in the plot above, are the differences between the *actual* (observed) *Y* values and the *Y* values that the regression equation predicts. These residuals represent variation in *Y* that the independent variable (X) does not account for (i.e. they represent the *error* in the model).

13.2.3 Formulas to calculate *a* and *b*

Fortunately, finding the equation of the line of best fit does not require summing the residuals of the infinite number of possible lines and selecting the line with smallest sum of squared residuals. Calculus provides simple equations for the intercept *b* and the slope *a* that minimize the SS of the residuals (i.e. the SSE):

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

For the sample dataset given above, we find:

$$b = \frac{\left[(4.6 - 4.98)(87.1 - 93.56) + \ldots + (5.1 - 4.98)(94.4 - 93.56)\right]}{(4.6 - 4.98)^2 + \ldots + (5.1 - 4.98)^2} = 7.69$$

$$a = 93.56 - 7.69(4.98) = 55.26$$

Therefore, the equation of the line of best fit is $Y = 55.26 + 7.69X$.

13.2.4 Covariance

In the formula for the slope given above, the quantity 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***, which is a quantity that indicates the degree to which the values of the two variables vary together. If high values of Y (relative to $\overline{Y}$) are associated with high values of X (relative to $\overline{X}$), the sample covariance will be positive. If high values of Y are associated with low values of X, or vice-versa, the sample covariance will be negative. If there is no association between the two variables, the sample covariance will be close to zero.

13.2.5 Using R for regression analysis

The lm() function can be used for regression analysis, as seen before when we discussed trend analysis. Representative code for the sample dataset above:

```
X <- c(4.6, 4.7, 5.1, 5.1, 4.8, 5.2, 4.4, 4.9, 5.9, 5.1)
Y <- c(87.1, 92.1, 93.1, 95.5, 89.8, 99.3, 91.4, 93.4, 99.5, 94.4)

regression <- lm(Y ~ X)
anova(regression)
summary(regression)
```

Output

```
Analysis of Variance Table
Response: Y
          Df   Sum Sq   Mean Sq  F value    Pr(>F)
X          1 90.83551 90.835510 16.23204 0.0037939 **
Residuals  8 44.76849  5.596061

Coefficients:
             Estimate Std. Error t value   Pr(>|t|)
(Intercept) 55.263281   9.534890  5.7959 0.00040706 ***
X            7.690104   1.908735  4.0289 0.00379385 **
Multiple R-squared:  0.6698586
```

This analysis tells us that the model accounts for a significant (p = 0.0038) amount of the variation in the experiment, nearly 67% of it (R-square = 0.67). This indicates that a great deal of the variation in food consumption among individuals is explained, through a simple linear relationship, by differences in body weight.

In the summary() output, we see that the equation of the best-fit line for this dataset is $Y = 55.26 + 7.69X$, just as we found before. The p-values associated with these estimates (0.0004 for a and 0.0038 for b) are the probabilities that the true values of these parameters are different from zero.

13.2.6 Analysis of adjusted *Y*'s

The experimental error in the previous analysis (MSE = 5.596) represents *the variation in food consumption that would have been observed if all the animals used in the experiment had had the same initial body weight*. To illustrate this, consider the following table in which each Y value is adjusted for differences in X via the regression equation. This adjustment essentially consists of sliding each value, in parallel with the best-fit line, to some common value of X. For this purpose, any value of X could be used to adjust the Y's, but $\overline{X}$ (4.98) is typically used as a representative value:

X	Y	Adjusted Y $= Y - b(X - \overline{X})$
4.6	87.1	90.0222
5.1	93.1	92.1772
4.8	89.8	91.1842
4.4	91.4	95.8602
5.9	99.5	92.4252
4.7	92.1	94.2532
5.1	95.5	94.5772
5.2	99.3	97.6082
4.9	93.4	94.0152
5.1	94.4	93.4772

The first adjusted value, 90.02224, is the food consumption expected for this animal *if its initial body weight had been 4.98 ($\overline{X}$)*. Because X and Y are positively correlated, the adjusted food consumption for underweight animals is always higher than the observed values and the adjusted food consumption for overweight animals is always lower.

Now consider the results of a regression on the adjusted Y's:

```
adjY <- Y - 7.69 * (X - mean(X))

adj_regression <- lm(adjY ~ X)
anova(adj_regression)
summary(adj_regression)
```

Output

```
Analysis of Variance Table

Response: adjY
          Df   Sum Sq   Mean Sq F value  Pr(>F)
X          1  0.00000 0.0000000       0 0.99996
Residuals  8 44.76849 5.5960612
```

```
Coefficients:
             Estimate   Std. Error t value   Pr(>|t|)
(Intercept) 9.355948e+01 9.534890e+00 9.81233 9.7761e-06 ***
X           1.041667e-04 1.908735e+00 0.00005    0.99996
Multiple R-squared:  3.722857e-10
```

As you might expect, with all animals adjusted to the same initial weight, body weight (X) no longer explains any variation in the study (SSX = 0, slope ~ 0). But here's the interesting result: The MSE (5.596) is exactly the same as we saw before! It is this uniformity in the MSE that shows that the two analyses are equivalent. That is, adjusting each Y to a common X by the best-fit equation is equivalent, in terms of accounting for variation, to a linear regression.

13.3 ANCOVA example

The analysis of covariance is illustrated below with data from a pilot experiment designed to study oyster growth. Specifically, the goals of this experiment were:

1. To determine if exposure to artificially-heated water affects growth
2. To determine if position in the water column (surface vs. bottom) affects growth

In this experiment, twenty bags of ten oysters each were placed across 5 locations within the cooling water runoff of a power-generation plant (i.e. 4 bags / location). Each location is considered a treatment: TRT1: cool-bottom, TRT2: cool-surface, TRT3: hot-bottom, TRT4: hot-surface, TRT5: control (i.e. mid-depth and mid-temperature).

Each bag of ten oysters is considered to be one experimental unit. The oysters were cleaned and weighed at the beginning of the experiment and then again about one month later. The dataset consists of the initial weight and final weight for each of the twenty bags.

The data:

Trtmt	Rep	Initial	Final
1	1	27.2	32.6
1	2	32	36.6
1	3	33	37.7
1	4	26.8	31
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

4	2	21.8	27
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

The code:

```
# I. Simple overall regression
oyster_reg_mod<-lm(Final ~ Initial, oyster_dat)
anova(oyster_reg_mod)
summary(oyster_reg_mod)

# II. Using loops in R to perform regressions at each treatment level
Trtmt_levels<-c(1:5)
for (i in Trtmt_levels) {
  with(subset(oyster_dat, Trtmt == Trtmt_levels[i]), {
    print(Trtmt_levels[i])
    print(summary(lm(Final ~ Initial)))
  })
}

# III. The one-way ANOVA
oyster_anova_mod<-lm(Final ~ Trtmt, oyster_dat)
anova(oyster_anova_mod)

# IV. The ANCOVA
#library(car)
oyster_ancova_mod<-lm(Final ~ Trtmt + Initial, oyster_dat)
anova(oyster_ancova_mod)
Anova(oyster_ancova_mod, type = 2)
summary(oyster_ancova_mod)
```

The first linear model performs a simple linear regression of Final Weight on Initial Weight and shows that, for the experiment as a whole, there is a significant linear relationship between these two variables ($p < 0.0001$; $R^2 = 0.95$), as shown in the output below:

Simple regression

```
Analysis of Variance Table

Response: Final
           Df    Sum Sq   Mean Sq   F value     Pr(>F)
Initial     1 342.35782 342.35782 377.79308 1.5761e-13 ***
Residuals 18  16.31168   0.90620
```

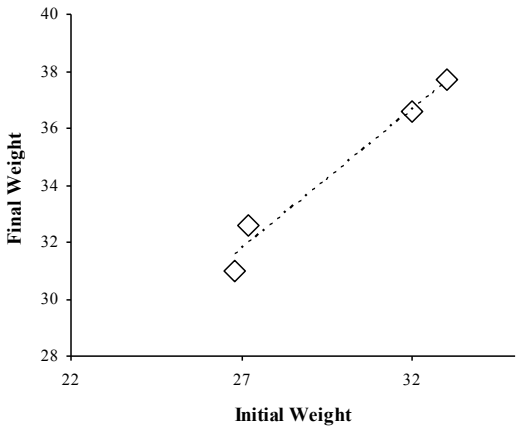
```
Coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 3.7646865  1.4094093  2.67111   0.015577 *
Initial     1.0512544  0.0540855 19.43690 1.5761e-13 ***
Multiple R-squared:  0.9545217
```

This strong dependence of Final Weight on Initial Weight suggests that Initial Weight may be a useful covariable for this analysis. The second part of the above script carries out a similar analysis *within each treatment group separately*. This analysis reveals the fact that the slope of this regression is fairly uniform across all treatment levels. This is important because in ANCOVA, all treatment groups are adjusted by the *same* slope. The estimates of the slopes within each treatment group:

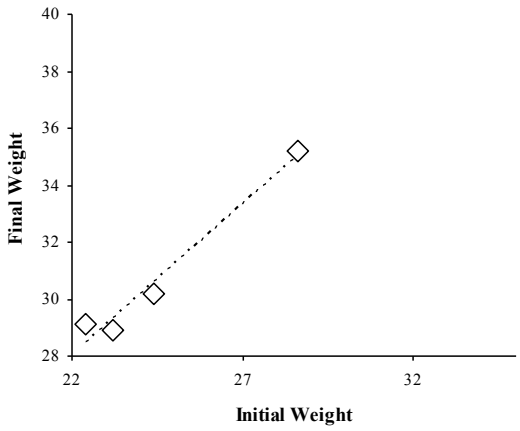
```
Coefficients:
                Estimate Std. Error  t value   Pr(>|t|)
Slope(Trt1)    0.9826468  0.1094183  8.98064   0.012173 *
Slope(Trt2)    1.5013550  0.3923086  3.82697   0.061997 .
Slope(Trt3)    1.0560666  0.1280121  8.24974   0.014377 *
Slope(Trt4)   1.05692503 0.06842649 15.44614  0.0041652 **
Slope(Trt5)   1.22388605 0.02530981 48.35619 0.00042738 ***
```

This similarity can be seen in the following scatterplots of Final vs. Initial Weight for treatment levels 1 and 3 below:

TRT1: $Y = 5.24 + 0.98X$



TRT3: $Y = 4.82 + 1.06X$



The third section of the above code conductrs a simple ANOVA of a CRD with four replications. The output:

The ANOVA

```
Response: Final
           Df   Sum Sq   Mean Sq F value   Pr(>F)
Trtmt       4 198.4070 49.601750 4.64255 0.012239 *
Residuals  15 160.2625 10.684167
```

From these results, we would conclude that location does affect oyster growth (p = 0.0122). This particular model explains roughly 55% of the observed variation (198.4 / (198.4 + 160.3)).

Finally, in the last section of the above code (the ANCOVA), we ask the question: What is the effect of location on Final Weight, adjusting for differences in Initial Weight? That is, what would the effect of Location be if all twenty bags of oysters had started with the same initial weight? The output:

The ANCOVA (Type I SS)

```
Analysis of Variance Table

Response: Final
           Df     Sum Sq    Mean Sq   F value     Pr(>F)
Trtmt       4 198.407000  49.601750 164.46503 1.3398e-11 ***
Initial     1 156.040177 156.040177 517.38400 1.8674e-12 ***
Residuals  14   4.222323   0.301595
```

The ANCOVA (Type II SS)

```
Anova Table (Type II tests)

Response: Final
              Sum Sq Df  F value     Pr(>F)
Trtmt      12.089359  4  10.0212 0.00048186 ***
Initial   156.040177  1 517.3840 1.8674e-12 ***
Residuals   4.222323 14
```

There are several things to notice here. First, the Type I SS for Trtmt (198.4) is the **unadjusted treatment SS** and is the same as the one found in the one-way ANOVA (previous page). 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 II SS for Trtmt (12.1) is the **adjusted treatment SS** and allows us to test the treatment effects, adjusting for other factors in the model. The reason adjustments are needed is because the two factors in the model, class variable Trtmt and regression variable INITIAL, are not orthogonal to one another. Because INITIAL is a continuous variable, the design is not balanced, even though there are no missing data (i.e. not all levels of Trtmt are present in all levels of Initial). This lack of orthogonality necessitates the use of partial sums of squares.

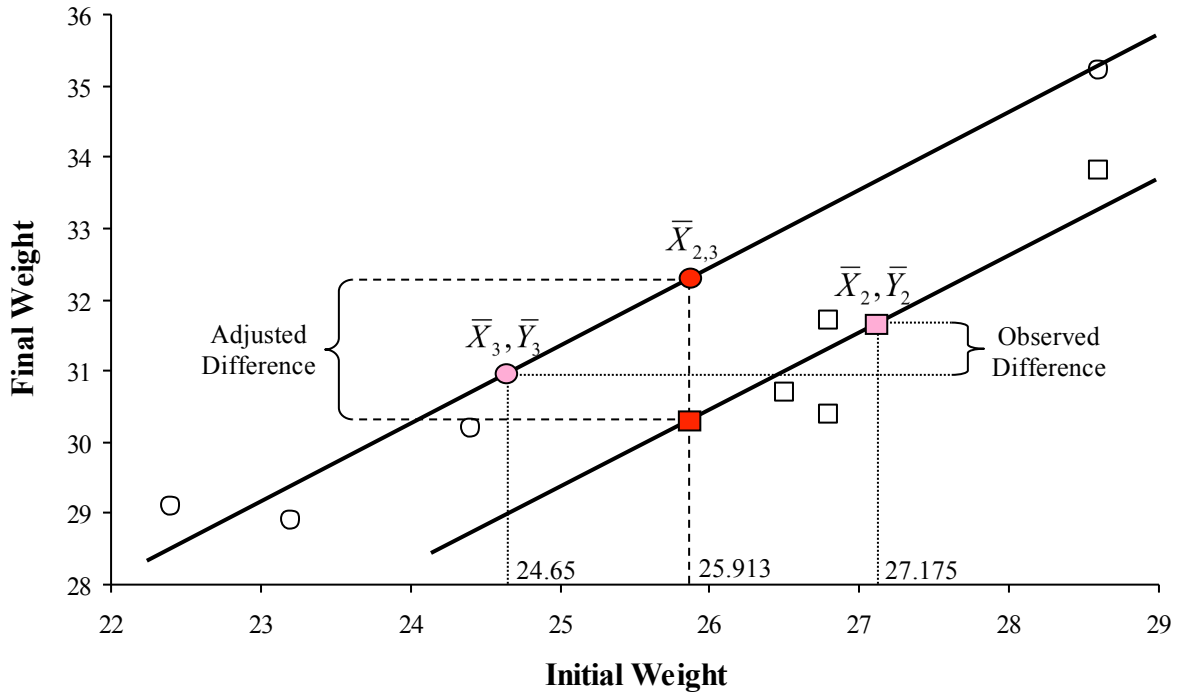
Type II SS produces the appropriate results in ANCOVA.

Even though the adjusted Trtmt MS (12.09/4 = 3.02) is much smaller than the unadjusted TRT MS (49.60), the reduction in the MSE is also quite large (from 10.68 in the ANOVA to 4.22/14 = 0.30 in the ANCOVA). It is this reduction in the experimental error that drives the increase in the F statistic for TRT from 4.64 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 among INITIAL values.

Similarly, the true test of the significance of the linear components of the relationship between INITIAL (X) and FINAL (Y) uses an INITIAL SS that is *adjusted for the effects of treatment*. In this case, notice that the INITIAL SS decreased (from 342.36 in the simple regression to 156.04 in the ANCOVA) because some of the observed variation can be attributed to the treatments. But again, the MSE also decreased significantly (from 0.91 to 0.30), ultimately leading to a more sensitive test for INITIAL.

13.3.1 Graphic interpretation of the ANCOVA example

The following scatterplot shows the data for treatments 2 (white squares) and 3 (white circles) from the oyster example. The mean final weight of treatment 3 (pink circle, 30.85) is seen to be slightly lower than the mean final weight of treatment 2 (pink square, 31.65).



For each treatment, variation in *X* is seen to contribute to variation in *Y*, as indicated by the common regression lines (solid lines). Because of this relationship, differences in the initial average weights of oysters assigned to each treatment can contribute greatly to the observed

differences between the final average weights. For example, Treatment 3 started with an initial average weight of 24.65, while Treatment 2 started with an initial average weight of 27.175. It is therefore likely that the final difference in weights (30.850 < 31.650) is not a good indicator of the treatment effects because the difference is due to *both* treatment effects *and* the differences in initial weights.

Thus the need to adjust the observed treatment means to some common initial weight. In the schematic above, this common initial weight is the mean of Treatments 2 and 3 (25.913). Adjustment consists of sliding the values of Treatment 3 up its regression line and the values of Treatment 2 down its regression line such that the initial weights of the two treatments are equal to the overall mean. By adjusting in this way, we see that the real effect of Treatment 3 is to increase the final weights of the oysters relative to Treatment 2. This effect was completely hidden in the original data.

13.3.2 Least squares adjusted means

If you requested the means of the treatment groups with a line like this:

```
Final_means <- aggregate(oyster_dat$Final,
      list(oyster_dat$Trtmt), mean)
```

R would produce the unadjusted means of the final weights of all five treatment levels. As discussed in the graphic example above, these means and the comparisons among them are not strictly appropriate.

To compare the true effects of the treatments, unbiased by differences in initial weights, the treatment means should be adjusted to what their values *would have been if all oysters had had the same initial weight*. As we saw before with unbalanced designs, these adjusted means can be calculated in R via the lsmeans() function. For example, the statement:

```
oyster.lsm <- lsmeans(oyster_ancova_mod, "Trtmt")
```

will generate an object containing least squares adjusted means that can then be acted upon by various means comparisons procedures via the contrast() function. In the summary table below, note the large differences between the unadjusted and adjusted treatments means for the variable FINAL:

TRT	Unadjusted Means	Adjusted LS Means	Calculation [$\bar{Y}_{adj_i} = \bar{Y}_i - \beta(\bar{X}_i - \bar{X})$]
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 differences are due to the large differences in initial weights among the treatment groups (TRT 5, for example, was assigned much smaller oysters than other treatments). In calculating these adjusted means, the coefficient $\beta = 1.08318$ is used. This coefficient can be found in the summary() output and represents a "best" single slope value that describes the relationship between X and Y, accounting for all other classification variables:

```
> summary(oyster_ancova_mod)

Call:
lm(formula = Final ~ Trtmt + Initial, data = oyster_dat)

Residuals:
       Min          1Q      Median          3Q         Max
-0.8438076 -0.3154120 -0.2170735  0.4863336  0.8871085

Coefficients:
               Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  2.25040039  1.44307538  1.55945 0.14120460
Trtmt2      -0.03581197  0.40722674 -0.08794 0.93116903
Trtmt3       1.89921708  0.45801799  4.14660 0.00098809 ***
Trtmt4       1.35157290  0.41936648  3.22289 0.00613476 **
Trtmt5       0.24445938  0.57658196  0.42398 0.67802248
Initial      1.08317982  0.04762051 22.74608 1.8674e-12 ***
```

This estimated "best" slope is identical to the slope one would obtain by performing individual ANOVAs on both X and Y, calculating their respective residuals, and then running a regression of the **Y residuals on the X residuals**. As suggested in the above table, this slope can be used to create a new adjusted response variable:

$$Z = Y - \beta(X - \bar{X})$$

This adjusted response variable (Z) is very handy because it can be used to perform a Levene's test for homogeneity of variances as well as a Tukey test for non-additivity, if the design is an RCBD with only one observation per block-treatment combination.

13.3.3 Contrasts

In this particular oyster example, the adjusted treatment means from the ANCOVA can be analyzed further with four orthogonal contrasts, as shown:

```
#Comparing LSMeans, using the "lsmeans" package (function contrast())
oyster.lsm <- lsmeans(oyster_ancova_mod, "Trtmt")

#Contrasts
contrast(oyster.lsm, list("control vs. trtmt"=c(-1,-1,-1,-1,4),
                          "bottom vs. surface"=c(-1,1,-1,1,0),
                          "cool vs. hot"=c(-1,-1,1,1,0),
                          "depth*temp"=c(1,-1,-1,1,0)))
```

The output:

contrast	estimate	SE	df	t.ratio	p.value
control.vs..trtmt	-2.2371404940	1.7037332311	14	-1.313	0.2103
bottom.vs..surface	-0.5834561450	0.5504960078	14	-1.060	0.3071
cool.vs..hot	3.2866019399	0.6157932555	14	5.337	0.0001
depth.temp	-0.5118322117	0.5869457568	14	-0.872	0.3979

The output indicates that oyster growth is only significantly affected by differences in temperature (cool vs. hot). Although constructed to be orthogonal, these contrasts are not orthogonal to the covariable; therefore, their sums of squares (if you were to go calculate them) do not add to the adjusted treatment SS.

Now consider this: If the covariable is **not** included in the model, these exact same contrasts produce completely different results:

```
> contrastmatrix<-cbind(c(-1,-1,-1,-1,4),c(-1,1,-1,1,0),
                        c(-1,-1,1,1,0),c(1,-1,-1,1,0))
> oyster_contrast_mod<-aov(Final ~ Trtmt, oyster_dat)
> summary(oyster_contrast_mod, split = list(Trtmt = list("Cont v. Trt" = 1,
      "Bot vs. Surf" = 2, "Cool vs. Hot" = 3, "Depth*Temp" = 4)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Trtmt	4	198.4070	49.60175	4.64255	0.012239	*
Trtmt: Cont v. Trt	1	169.3620	169.36200	15.85168	0.001204	**
Trtmt: Bot vs. Surf	1	2.1025	2.10250	0.19679	0.663659	
Trtmt: Cool vs. Hot	1	9.3025	9.30250	0.87068	0.365545	
Trtmt: Depth*Temp	1	17.6400	17.64000	1.65104	0.218305	
Residuals	15	160.2625	10.68417			

13.4 ANCOVA model

Recall the ANOVA model for a CRD:

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

We have not discussed the linear model for a simple linear regression, but it is:

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

ANCOVA is a combination of ANOVA and regression, a fact reflected in its linear model:

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

Extending this concept, the linear model for ANOCOVA within any given design (e.g. CRD, RCBD, LS, etc.) is simply the linear model for the ANOVA *plus* an additional term for the concomitant variable. For the CRD, the formula can be slightly rearranged:

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

And here you have it: An ANCOVA on the original response variable (Y) is equivalent to a regular ANOVA on values of Y that have been adjusted according to their linear dependence on X. In the discussion which follows (Topic 13, Part II), we denote these regression-adjusted values with the letter Z.