Lecture 19

Topic 13: Analysis of Covariance (ANCOVA), Part II

Assumptions of ANCOVA

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

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

fixed: inferences are appropriate for interpolated rather than extrapolated values

measured without error: measurement error is trivial relative to observed variation

independent of treatments: the treatments themselves do not affect the X values

3. The regression of Y on X is linear and independent of treatments.

linear: approximately linear within the given range of X

independent of treatments: regression slopes are homogeneous across treatment levels

X values are independent of the treatments

Was the covariable measured **before** or **after** applying the treatments?

Before: Independent, by definition.
After: Independence should be investigated.

An ANOVA of the *covariable* (X) is appropriate to test this hypothesis.

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

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

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
```

```
> mean(oyster_dat$Initial)
[1] 25.76
```

ANCOVA can be used where the *X* values are affected by the treatments, but results should be interpreted with caution.

Slopes are homogeneous across treatment levels

Because a single slope is used to adjust all observations in the experiment, the covariate coefficients must be the same for each level of the categorical variable being analyzed.

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:**

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

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
```

p = 0.3602 > 0.05, so we fail to reject the null hypothesis that slopes are homogeneous across treatment levels

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

Homogeneity of variances

PROBLEM: Levene's Test is only defined for **one-way ANOVAs**.

```
# The ANCOVA
lm(Final ~ Trtmt + Initial, oyster_dat)    ← Not a one-way ANOVA!
```

SOLUTION: Manually adjust the response variable (Y → Z), where

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

```
# The equivalent ANCOVA
lm(Z ~ Trtmt, oyster_dat)    ← Voila!  A one-way ANOVA!
leveneTest(Z ~ Trtmt, data = oyster_dat)
```

Normality of residuals

The usual approach is followed:

```
#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)

#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

Additivity of main effects

In an RCBD with one replication per block-treatment combination, the adjusted Z values should also be used for the Tukey 1-df Test for Non-additivity:

```
#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)
```

The full analysis:

```
#Inform R about which variables are factors
oyster_dat$Trtmt<-as.factor(oyster_dat$Trtmt)
oyster_dat$Rep<-as.factor(oyster_dat$Rep)

# 1. General regression
oyster_reg_mod<-lm(Final ~ Initial, oyster_dat)
anova(oyster_reg_mod)

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

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

# 4.  Perform ANOVAs for both X and Y
oyster_anovaX_mod<-lm(Initial ~ Trtmt, oyster_dat)
anova(oyster_anovaX_mod)

oyster_anovaY_mod<-lm(Final ~ Trtmt, oyster_dat)
anova(oyster_anovaY_mod)

# 5.  Test standard ANOVA assumptions, using Z
# a.  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)

# b.  Homogeneity of variances
#library(car)
leveneTest(Z ~ Trtmt, data = oyster_dat)

# 6.  Test ANCOVA assumptions
# a.  Homogeneity of slopes
oyster_slopes_mod<-lm(Final ~ Trtmt + Initial + Trtmt:Initial, oyster_dat)
anova(oyster_slopes_mod)

# 7.  The ANCOVA, with desired subsequent analysis
#library(car)
oyster_ancova_mod<-lm(Final ~ Trtmt + Initial, oyster_dat)
Anova(oyster_ancova_mod, type = 2)
summary(oyster_ancova_mod)

#Compare LSMeans, using the "lsmeans" package (function contrast())
oyster_lsm <- lsmeans(oyster_ancova_mod, "Trtmt")
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)))
```

Relative efficiency

$$RE_{1:2} = \frac{I_1}{I_2} = \frac{MSE_2}{MSE_1}$$

In the oyster example:

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

The **effective ANCOVA MSE**, adjusting for sampling error in X:

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

An estimate of the relative precision:

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

Each replication, adjusted for the effect of the covariable, is as effective as **26.6 replications** without such adjustment.

Uses of ANCOVA

The most important uses of the analysis of covariance are:

1. To **control error** and thereby 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 <u>interpretation of data</u>**, especially with regard to the nature or mechanism of treatment effects.

Interpretation of ANCOVA examples

1. **Effect of a fertilizer treatment on sugar beets.**

Y = Yield X = Number of plants per plot

	Treatment	Treatment	Treatment
ANOVA	Significant	Significant	NS
ANCOVA	Significant	NS	Significant

2. **Effect of different chicken feeds on weight gain.**

Y = Gained weight X = Food intake

	Treatment	Treatment	Treatment
ANOVA	Significant	Significant	NS
ANCOVA	Significant	NS	Significant

3. **Effect of two storage protein genes on bread loaf volume (2x2 factorial).**

Y = Loaf volume X = Grain protein content

	Treatment	Treatment	Treatment
ANOVA	G1 = S, G2 = S	G1 = NS, G2 = NS	G1 = S, G2 = S
ANCOVA	G1 = NS, G2 = NS	G1 = S, G2 = S	G1 = NS, G2 = S

4. **Effect of grain hardness gene and environment on grain hardness.**

Y = Grain hardness X = Grain weight

	Treatment		Treatment	
ANOVA	G = S, E = S	GxE = S	G = NS, E = S	GxE = S
ANCOVA	G = S, E = NS	GxE = NS	G = S, E = NS	GxE = S

Free Bonus! R Coding Extravaganza

ANCOVA for a 3x2 Factorial arranged as an RCBD

1. General regression

```
reg_mod<-lm(Y ~ X, fact_dat)
anova(reg_mod)
```

2. Find beta and Xmean

```
ancova_mod<-lm(Y ~ Block + A + B + A:B + X, fact_dat)
summary(ancova_mod)
mean(fact_dat$X)
```

3. Create Z

```
fact_dat$Z<-fact_dat$Y - beta*(fact_dat$X - Xmean)
```

4. Perform ANOVA on Y

```
anovaY_mod<-lm(Y ~ Block + A + B + A:B, fact_dat)
anova(anovaY_mod)
```

5. Test standard ANOVA assumptions, using Z

```
# Normality of residuals (ANOVA on Z)
anovaZ_mod<-lm(Z ~ Block + A + B + A:B, fact_dat)
fact_dat$anovaZ_resids <- residuals(anovaZ_mod)
shapiro.test(fact_dat$anovaZ_resids)
```

For Levene's, create a Treatment ID (TRT) with 6 levels, one for each A-B combination:

```
# Homogeneity of variances
#library(car)
leveneTest(Z ~ Trtmt, data = fact_dat)

# Additivity of main effects
fact_dat$anovaZ_preds <- predict(anovaZ_mod)
fact_dat$sq_anovaZ_preds <- fact_dat$anovaZ_preds^2
tukeyZ_mod<-lm(Z ~ Block + A + B + A:B + sq_anovaZ_preds, fact_dat)
anova(tukeyZ_mod)
```

6. Test ANCOVA assumptions

```
# Homogeneity of slopes
slopes_mod<-lm(Y ~ Block + Trtmt + X + Trtmt:X, fact_dat)
anova(slopes_mod)

# Independence of X from treatments
anovaX_mod<-lm(X ~ Block + A + B + A:B, fact_dat)
anova(anovaX_mod)
```

7. The ANCOVA

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

8. To partition the A:B interaction

```
fact_lsm <- lsmeans(ancova_mod, "Trtmt")
contrast(fact_lsm, list(  "A lin"       =c(-1,0,1,-1,0,1),
                          "A quad"      =c(1,-2,1,1,-2,1),
                          "B"           =c(1,1,1,-1,-1,-1),
                          "A lin * B"   =c(-1,0,1,1,0,-1),
                          "A quad * B"  =c(1,-2,1,-1,2,-1)  ) )
```

9. Trend analysis on A (unequally-spaced treatment levels)

Re-import the dataset, leaving A as a numeric regression variable

```
A<-fact_dat$A
A2<-A^2

fact_trend_mod<-(lm(Y ~ X + Block + B + A + A2 + A:B + A2:B, fact_dat))
anova(fact_trend_mod)
```

10. To analyze a mixed model (Factor A random, Factor B fixed)

```
What do you do?
```