

Stat 141 (Experimental Designs I)

Analysis of Covariance

Quick review

- **Nuisance factor**- factor that may have an effect on the response variable but of not an interest to the researcher
 - *Unknown and uncontrollable*- randomization balances out its impact across the experiment
 - *Known and controllable*- blocking or grouping of the experimental units using the nuisance factor as grouping variable
 - *Known but uncontrollable*- measure the value of this factor and include them in the analysis using **Analysis of Covariance**

Some examples

1. An experiment designed to test the effects of three diets on yearling weight of animals, different initial weight or different age at the beginning of the experiment will influence the precision of the experiment. It is necessary to adjust yearling weights for differences in initial weight or initial age. This can be accomplished by defining initial weight or age as a covariate in the model. This will improve the precision of the experiment, since part of the unexplained variability is explained by the covariate and consequently the experimental error is reduced.
 2. A study is designed to evaluate different methods of teaching reading to 8-year-old children. The response variable is final scores of the children after participating in the reading program. However, the children participating in the study will have different reading ability prior to entering the program. Also, there will be many factors outside the school that may have an influence on the reading score of a child, such as socioeconomic variables associated with the child's family.
- The variables that describe the differences in experimental units or experimental conditions prior to the experiment are called *concomitant variables* or simply **covariates**

Analysis of covariance

- Method of adjusting the analysis for nuisance factors that cannot be controlled and that sometimes cannot be measured until the experiment is conducted
- The method is applicable if the nuisance factors are related to the response variable but are themselves unaffected by the treatment factors
- Combines ANOVA and regression analysis
- Two main uses:
 - To reduce the error variance by eliminating the unit-to-unit variation attributable to fluctuations in the covariates
 - To eliminate any bias in treatment comparisons caused by the uneven covariates assigned to the various experimental units by adjusting the treatment means

Analysis of covariance (CRD)

Linear Model:

$$Y_{ij} = \mu + \tau_i + \beta x_{ij} + \epsilon_{ij}$$

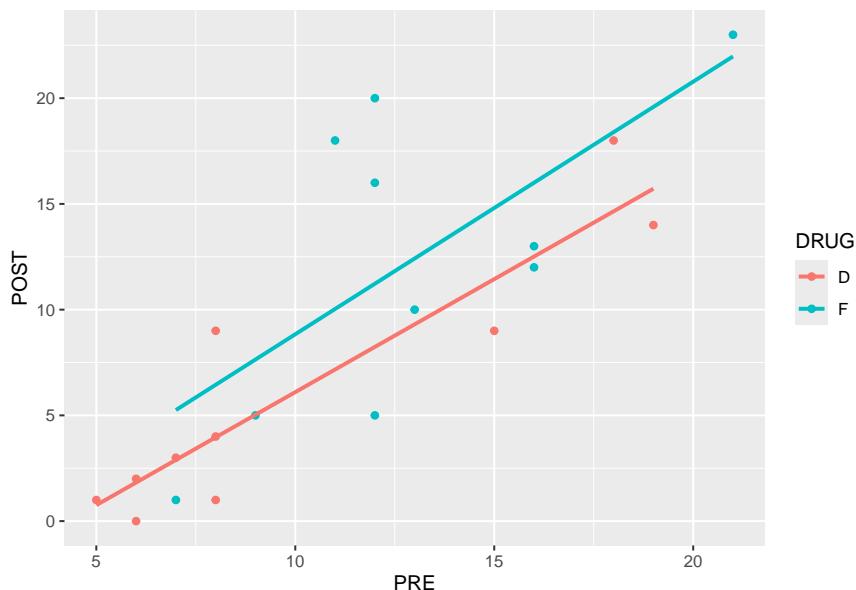
- the terms μ , τ_i , and ϵ_{ij} are as defined before
- x_{ij} is the value of the covariate
- β is the regression coefficient of X, defined as the expected unit change in Y for every unit increase in X
- Alternatively, we can use the following model:

$$Y_{ij} = \mu + \tau_i + \beta (x_{ij} - \bar{x}_{..}) + \epsilon_{ij}$$

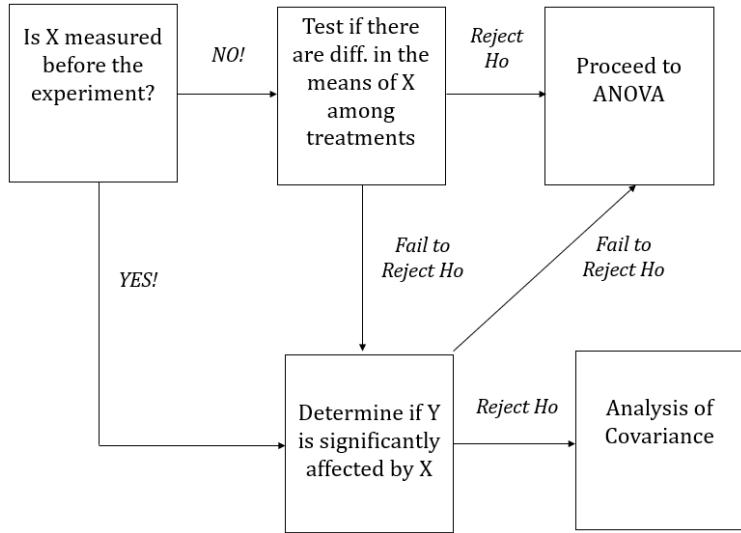
- The two models are equivalent for comparison of treatment effects
- The slope parameter β has the same interpretation in both models
- In the first model, $\mu + \tau_i$ is the mean response when $x_{ij} = 0$, whereas, in the alternative model, $\mu + \tau_i$ is the mean response when $x_{ij} = \bar{x}_{..}$
- The second model is usually preferred to reduce computational problems and is a little easier to work with in obtaining least squares estimates

Assumptions

- For both models, $\epsilon_{ij} \sim NID(0, \sigma^2)$
- The X's are fixed or measured without error or negligible error and **independent** of the treatments
- The regression function relating Y to the X's after removal of treatment effects is known, say **linear**
- The regression of Y on X is independent of the treatments and the X's are not affected by the treatments (**equal regression slopes**)



Process flow in analysis of covariance



1. Test if the covariate (X) is important
 - Assuming heterogeneous slopes, test $H_0 : \beta_1 = \beta_2 = \dots = \beta_t = 0$
 - Fit a regression model of this form: $Y = Trt + Trt : X$
 - Assuming homogeneous slopes, test $H_0 : \beta = 0$
 - Fit a regression model of this form: $Y = Constant + Trt + X$

REMARKS:

- If the term $TRT:X$ in the first test and the term X in the second test are BOTH non-significant, then X is not important. Drop X and proceed to ANOVA.
 - If either is significant, proceed to ANCOVA. Go to Step 2.
2. Test whether there is equal slope
 - We test $H_0 : \beta_1 = \beta_2 = \dots = \beta_t$
 - Fit a model of the form $Y = TRT + X + TRT : X$
 - If the term $TRT:X$ is significant, then follow a Johnson-Neyman analysis (*Reading assignment*). Otherwise, perform ANCOVA.
 3. Test of (adjusted) treatment effect
 - $H_0 : \tau_i = 0, \forall i$
 - Fit a model of the form: $Y = Constant + TRT + X$
 - Use *emmeans()* for post hoc analysis, if necessary

ANCOVA in CRD: an example

An animal scientist is interested in determining the effects of four different feed plans on hogs. Twenty four hogs of a breed were chosen and randomly assigned to one of the four feeding plans for certain period. Initial weight (X) of the hogs and gains in weight (Y) in pounds at the end of the experiment were recorded and stored in the file *Ancova_crd1.csv*. Since differences in initial weight may contribute to the differences in gain in weight, it is decided to use initial weight as a covariate in an analysis of covariance.

```
#Load data
ancov1 <- read.csv("Ancova_crd1.csv")

#Fit a model without intercept
mod1.1 <- lm(Wtgain ~ -1 + Feeds + Feeds:InitialWt,
              data = ancov1)

#Generate Type III sums of squares using Anova() function in car package
Anova(mod1.1, type=3)

## Anova Table (Type III tests)
##
## Response: Wtgain
##             Sum Sq Df F value    Pr(>F)
## Feeds          9087.8  4  8.6625 0.0006418 ***
## Feeds:InitialWt 1741.4  4  1.6599 0.2082316
## Residuals      4196.4 16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Fit another model (with intercept)
mod1.2 <- lm(Wtgain ~ Feeds + InitialWt,
              data=ancov1)

#Generate Type III sums of squares using Anova() function in car package
Anova(mod1.2, type=3)

## Anova Table (Type III tests)
##
## Response: Wtgain
##             Sum Sq Df F value    Pr(>F)
## (Intercept) 6731.8  1 24.3394 9.225e-05 ***
## Feeds        1609.6  3  1.9399   0.1574
## InitialWt     682.8  1  2.4688   0.1326
## Residuals     5255.0 19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

INTERPRETATION:

Since the term *Feeds:InitialWt* in the first model (**mod1.1**) and the term *InitialWt* in the second model (**mod1.2**) are both NOT significant, then the covariate (*InitialWt*) is not important and can be dropped in the analysis.

```

mod1.3 <- lm(Wtgain ~ Feeds, data=ancov1)
Anova(mod1.3, type=3)

## Anova Table (Type III tests)
##
## Response: Wtgain
##           Sum Sq Df F value Pr(>F)
## (Intercept) 146954  1 494.9735 1.4e-15 ***
## Feeds       2163   3  2.4286 0.09531 .
## Residuals    5938  20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

INTERPRETATION:

At $\alpha = 0.05$, the mean weight of hogs fed with four feed plans are not significantly different. In other words, different feed plans do not significantly affect weight gain.

Analysis of covariance in CRD: another example

Three antibiotics (A, D and F) are to be tested on leprosy patients. Ten patients are selected for each treatment (drug) and six sites on each patient are measured for leprosy bacili. The ultimate objective is determining the effect of the drugs on the post-treatment count of bacili. However, it is suspected that including the pre-treatment bacili count may improve the precision of the experiment. The data is stored in *Ancova_crd2.csv*.

```

#Reading the data into R
ancov2 <- read.csv("Ancova_crd2.csv")

#Fitting a model without an intercept (assuming unequal slopes)
mod2.1 <- lm(POST ~ -1 + DRUG + DRUG:PRE,
             data = ancov2)
Anova(mod2.1, type=3)

```

```

## Anova Table (Type III tests)
##
## Response: POST
##           Sum Sq Df F value Pr(>F)
## DRUG        54.40   3  1.0947  0.3705
## DRUG:PRE  597.54   3 12.0243 5.292e-05 ***
## Residuals 397.56  24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

#Fitting a model with an intercept (assuming equal slopes)
mod2.2 <- lm(POST ~ DRUG + PRE,
             data = ancov2)
Anova(mod2.2, type=3)

```

```

## Anova Table (Type III tests)
##

```

```

## Response: POST
##           Sum Sq Df F value    Pr(>F)
## (Intercept) 61.26  1 3.8177   0.06155 .
## DRUG        68.55  2 2.1361   0.13838
## PRE         577.90  1 36.0145 2.454e-06 ***
## Residuals   417.20 26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1

```

INTERPRETATION:

In model *mod2.1* DRUG:PRE is significant and in *mod2.2* PRE is significant, therefore X is important. We proceed to Step 2.

```

#Test for homogeneous slopes
mod2.3 <- lm(POST ~ DRUG + PRE + DRUG:PRE,
             data = ancov2)
Anova(mod2.3, type=3)

```

```

## Anova Table (Type III tests)
##
## Response: POST
##           Sum Sq Df F value    Pr(>F)
## (Intercept) 5.08   1 0.3064  0.58499
## DRUG        8.50   2 0.2566  0.77574
## PRE         113.35  1 6.8427  0.01515 *
## DRUG:PRE    19.64  2 0.5930  0.56058
## Residuals   397.56 24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1

```

INTERPRETATION:

The term DRUG:PRE is not significant, therefore the equal slopes assumption is met. We can proceed to ANCOVA.

```

mod2.4 <- lm(POST ~ DRUG + PRE,
             data = ancov2)
Anova(mod2.4, type=3)

```

```

## Anova Table (Type III tests)
##
## Response: POST
##           Sum Sq Df F value    Pr(>F)
## (Intercept) 61.26  1 3.8177   0.06155 .
## DRUG        68.55  2 2.1361   0.13838
## PRE         577.90  1 36.0145 2.454e-06 ***
## Residuals   417.20 26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1

```

INTERPRETATION:

There is no significant difference in the mean post-treatment count of *bacili* at $\alpha = 0.05$.

Analysis of covariance in RCBD

- Linear model: $Y_{ij} = \mu + \gamma_j + \tau_i + \beta(x_{ij} - \bar{x}_{..}) + \epsilon_{ij}$
- The analysis strategy for ANCOVA in RCBD is similar to that of ANCOVA in CRD.
- Just insert the block effect term in the code in R

ANCOVA in RCBD: an example

A varietal trial was conducted with 8 varieties (V_1, V_2, \dots, V_8) of hybrid maize. The trial was laid out in a randomized block design with 4 replications. At the time of harvest the number of plants per plot of size 48ftx12ft was also recorded along with the plot yield. The data of grain yield (Y) in lbs per plot and the number of plants (X) are contained in *Ancova_rbd.csv*.

```
#Load the data
ancov3 <- read.csv("Ancova_rbd.csv")
ancov3$Block <- factor(ancov3$Rep)

#Fit a model without intercept
mod3.1 <- lm(Y ~ -1 + Block + Variety + Variety:X,
               data=ancov3)

#Generate Type III sums of squares using Anova() function in car package
Anova(mod3.1, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: Y
##             Sum Sq Df F value    Pr(>F)
## Block      2383.62  4  2.7826 0.07175 .
## Variety     501.69  7  0.3347 0.92388
## Variety:X 1674.85  8  0.9776 0.49348
## Residuals  2783.98 13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod3.2 <- lm(Y ~ Block + Variety + X,
               data=ancov3)
Anova(mod3.2, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: Y
##             Sum Sq Df F value    Pr(>F)
## (Intercept)   1.8   1  0.0107 0.918807
## Block        2314.9   3  4.6088 0.013120 *
## Variety       4890.6   7  4.1729 0.005536 **
## X            1110.3   1  6.6313 0.018072 *
## Residuals    3348.6  20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

INTERPRETATION:

Since the covariate X in **mod3.2** is significant, then X is important and we must proceed to the next test.

```
mod3.3 <- lm(Y ~ Block + Variety + X +Variety:X,  
               data=ancov3)  
Anova(mod3.3, type=3)
```

```
## Anova Table (Type III tests)  
##  
## Response: Y  
##           Sum Sq Df F value    Pr(>F)  
## (Intercept) 18.72  1  0.0874 0.77214  
## Block       2294.53  3  3.5715 0.04417 *  
## Variety      501.69  7  0.3347 0.92388  
## X            571.10  1  2.6668 0.12644  
## Variety:X    564.59  7  0.3766 0.90006  
## Residuals   2783.98 13  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

INTERPRETATION:

Since the covariate the term *Variety:X* is not significant, the homogeneous slopes assumption is met and we can proceed to ANCOVA.

```
mod3.4 <- lm(Y ~ Variety + Block + X,  
               data = ancov3)  
Anova(mod3.4, type=3)
```

```
## Anova Table (Type III tests)  
##  
## Response: Y  
##           Sum Sq Df F value    Pr(>F)  
## (Intercept) 1.8  1  0.0107 0.918807  
## Variety     4890.6  7  4.1729 0.005536 **  
## Block       2314.9  3  4.6088 0.013120 *  
## X           1110.3  1  6.6313 0.018072 *  
## Residuals   3348.6 20  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

INTERPRETATION:

- There is significant differences in the mean yield per plot after adjusting for the blocks and the number of plants per plot.
- Proceed to post hoc analysis of the variety effect

```
emmeans(mod3.4, pairwise ~ Variety,  
        adjust = "tukey")
```

```

## $emmeans
##   Variety emmean    SE df lower.CL upper.CL
##   V1        52.9 6.79 20     38.7    67.0
##   V2        81.1 6.50 20     67.5    94.6
##   V3        47.1 6.74 20     33.0    61.2
##   V4        47.8 6.71 20     33.8    61.7
##   V5        52.0 6.54 20     38.4    65.7
##   V6        59.9 6.67 20     46.0    73.9
##   V7        52.7 6.50 20     39.1    66.2
##   V8        33.6 6.48 20     20.0    47.1
##
## Results are averaged over the levels of: Block
## Confidence level used: 0.95
##
## $contrasts
##   contrast estimate    SE df t.ratio p.value
##   V1 - V2   -28.220 9.26 20  -3.047  0.0950
##   V1 - V3    5.760 9.97 20   0.578  0.9988
##   V1 - V4    5.119 9.91 20   0.516  0.9994
##   V1 - V5    0.822 9.62 20   0.085  1.0000
##   V1 - V6   -7.070 9.16 20  -0.772  0.9928
##   V1 - V7    0.185 9.26 20   0.020  1.0000
##   V1 - V8   19.317 9.46 20   2.043  0.4802
##   V2 - V3   33.980 9.49 20   3.581  0.0324
##   V2 - V4   33.339 9.46 20   3.526  0.0363
##   V2 - V5   29.042 9.28 20   3.130  0.0810
##   V2 - V6   21.150 9.21 20   2.297  0.3416
##   V2 - V7   28.405 9.15 20   3.105  0.0850
##   V2 - V8   47.537 9.20 20   5.168  0.0010
##   V3 - V4   -0.641 9.15 20  -0.070  1.0000
##   V3 - V5   -4.939 9.20 20  -0.537  0.9993
##   V3 - V6  -12.830 9.81 20  -1.308  0.8853
##   V3 - V7   -5.575 9.49 20  -0.588  0.9987
##   V3 - V8   13.556 9.28 20   1.460  0.8186
##   V4 - V5   -4.298 9.19 20  -0.468  0.9997
##   V4 - V6  -12.189 9.76 20  -1.249  0.9069
##   V4 - V7   -4.934 9.46 20  -0.522  0.9994
##   V4 - V8   14.198 9.26 20   1.533  0.7816
##   V5 - V6   -7.892 9.50 20  -0.831  0.9890
##   V5 - V7   -0.637 9.28 20  -0.069  1.0000
##   V5 - V8   18.495 9.17 20   2.017  0.4955
##   V6 - V7    7.255 9.21 20   0.788  0.9919
##   V6 - V8   26.387 9.36 20   2.820  0.1451
##   V7 - V8   19.132 9.20 20   2.080  0.4588
##
## Results are averaged over the levels of: Block
## P value adjustment: tukey method for comparing a family of 8 estimates

```

Table of Means

Variety	Mean
V1	52.87 ^{ab}
V2	82.09 ^a
V3	47.11 ^b
V4	47.75 ^b
V5	52.05 ^{ab}
V6	59.94 ^{ab}
V7	52.69 ^{ab}
V8	33.55 ^b