Analysis of Covariance (ANCOVA)

In some experiments, the experimental units are nonhomogeneous or there is a variation in the experimental conditions that are not due to the treatments but might have an effect on the response variable. The *quantitative* variables that describe the differences in experimental units or experimental conditions are called ***covariates*** or ***concomitant*** variables. The analysis of covariance (ANCOVA) is a method by which the influence of the covariates on the treatment means is reduced. This will often result in increased precision for parameter estimates and increased power for tests of hypothesis. ANCOVA combines features of analysis of variance and regression and can be used for either observational studies or designed experiments.

- **Examples**

1. A company studied the effects of three different types of promotion on sales of its crackers:

- Treatment 1 - Sampling of product by customers in store and regular shelf space
- Treatment 2 - Additional shelf space in regular location
- Treatment 3 - Special display shelves at ends of aisle in addition to regular shelf space

Fifteen stores were selected for the study, and five randomly selected stores were assigned to each treatment. Other relevant conditions under the control of the company, such as price and advertising, were kept the same for all stores in the study. Data on the number of cases of the product sold during the promotional period (Y) and data on sales of the product in the preceding period (X) are given in the table below.

Trt	Y_{i1}	X_{i1}	Y_{i2}	X_{i2}	Y_{i3}	X_{i3}	Y_{i4}	X_{i4}	Y_{i5}	X_{i5}
1	38	21	39	26	36	22	45	28	33	19
2	43	34	38	26	38	29	27	18	34	25
3	24	23	32	29	31	30	21	16	28	29

2. An experiment was conducted to see the effects of two treatments, a slow-release fertilizer (s) and a fast-release fertilizer (f), on seed yield (grams) of peanut plants were compared with a control (c), a standard fertilizer. Ten replications of each treatment were to be grown in a greenhouse study. When setting up the experiment, the researcher recognized that the 30 peanut plants were not exactly at the same level of development or health. Consequently, the researcher recorded the height (cm) of the plant, a measure of plant development and health, at the start of the experiment. The results of the experiment are shown in the table below.

c		s		f	
Yield	Height	Yield	Height	Yield	Height
12.2	45	16.6	63	9.5	52
12.4	52	15.8	50	9.5	54
11.9	42	16.5	63	9.6	58
11.3	35	15.0	33	8.8	45
11.8	40	15.4	38	9.5	57
12.1	48	15.6	45	9.8	62
13.1	60	15.8	50	9.1	52
12.7	61	15.8	48	10.3	67
12.4	50	16.0	50	9.5	55
11.4	33	15.8	49	8.5	40

- **Concomitant Variables/Covariates**

  1. If the concomitant variable has no relation to the response variable, nothing is gained by covariance analysis.

  2. Covariates should not be influenced by the treatments in any way.
  A company was conducting a training school for engineers to teach them accounting and budgeting principles. Two teaching methods (computer-assisted learning and standard lecture) were used, and engineers were assigned at random to one of the two. At the end of the program, a score was obtained for each engineer reflecting the amount of learning. The analyst decided to use as a covariate in covariance analysis the amount of time devoted to study (which the engineers were required to record). After conducting the analysis of covariance, the analyst found that training method had virtually no effect.

- **Covariance Model**

$$Y_{ij} = \mu_{.} + \tau_i + \gamma(X_{ij} - \bar{X}_{..}) + \epsilon_{ij}$$

  where:

  1. $\mu_{.}$ is an overall mean.

  2. τ_i are the fixed treatment effects, subject to the restriction $\sum \tau_i = 0$.

  3. γ is a regression coefficient for the relation between Y and X.

  4. X are constants

  5. ϵ_{ij} are independent $N(0, \sigma^2)$.

- **Review: R Commands for Linear Regression**

```
# Tree circumference and height problem
> circ=c(1.8,1.9,1.8,2.4,5.1,3.1,5.5,5.1,8.3,13.7,5.3,4.9,3.7,3.8)
> height=c(21,33.5,24.6,40.7,73.2,24.9,40.4,45.3,53.5,93.8,64,62.7,47.2,44.3)

> plot(circ,height,xlab="Tree Circumference",ylab="Tree Height",main="Scatter Plot")
> cor(circ,height)            # Computers the correlation coefficient
[1] 0.8283738

> result=lm(height˜circ)
> summary(result)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   22.463      5.875   3.824 0.002424 **
circ           5.341      1.043   5.123 0.000252 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.87 on 12 degrees of freedom
Multiple R-squared: 0.6862,     Adjusted R-squared: 0.6601
F-statistic: 26.24 on 1 and 12 DF,  p-value: 0.0002522

> abline(result)             # Plots the regression line over the scatterplot

> qqnorm(result$residuals); shapiro.test(result$residuals)
> plot(result$fitted,result$residuals)

> confint(result)             # Confidence intervals for slope and y-intercept parameters
               2.5 %    97.5 %
(Intercept) 9.662851 35.263001
circ        3.069104  7.612192

> new=data.frame(circ=c(10,12))
# Predicts the mean value of Y at X=10 and X=12.
> predict(result,newdata=new,interval="confidence",level=.90)
      fit      lwr      upr
1 75.8694 64.58271  87.1561
2 86.5507 71.92861 101.1728
```

```
# Predicts a value for Y at X=10 and X=12.
> predict(result,newdata=new,interval="prediction",level=.90)
      fit      lwr       upr
1 75.8694 51.89217  99.84664
2 86.5507 60.83449 112.26692
```

- **R Commands for ANCOVA Example**

```
> data=read.csv("Crackers.csv",header=T)
> attach(data)

> plot(treatment,new_sales)

> plot(past_sales,new_sales,pch=treatment)
> legend(16,45,c("Trt1","Trt2","Trt3"),pch=c(1,2,3))

> anova(aov(past_sales~factor(treatment)))
Response: past_sales
                  Df Sum Sq Mean Sq F value Pr(>F)
factor(treatment)  2   26.8  13.400  0.4826 0.6287
Residuals         12  333.2  27.767

> anova(aov(new_sales~factor(treatment)))
Response: new_sales
                  Df Sum Sq Mean Sq F value  Pr(>F)
factor(treatment)  2  338.8 169.400  6.6086 0.01161 *
Residuals         12  307.6  25.633

> result1=lm(new_sales~past_sales+factor(treatment))
> anova(result1)
Response: new_sales
                  Df Sum Sq Mean Sq F value    Pr(>F)
past_sales         1 190.68 190.678  54.379 1.405e-05 ***
factor(treatment)  2 417.15 208.575  59.483 1.264e-06 ***
Residuals         11  38.57   3.506

> summary(result1)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         17.3534     2.5230   6.878 2.66e-05 ***
past_sales           0.8986     0.1026   8.759 2.73e-06 ***
factor(treatment)2  -5.0754     1.2290  -4.130  0.00167 **
factor(treatment)3 -12.9768     1.2056 -10.764 3.53e-07 ***

> past2=past_sales-mean(past_sales)
> result2=lm(new_sales~past2+factor(treatment))
> summary(result2)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         39.8174     0.8576  46.432 5.66e-14 ***
past2                0.8986     0.1026   8.759 2.73e-06 ***
factor(treatment)2  -5.0754     1.2290  -4.130  0.00167 **
factor(treatment)3 -12.9768     1.2056 -10.764 3.53e-07 ***
```