ANCOVA

Analysis of Covariance

History

Fisher introduced "analysis of covariance" in "Studies in Crop Variation. IV“ (Eden and Fisher. 1927).

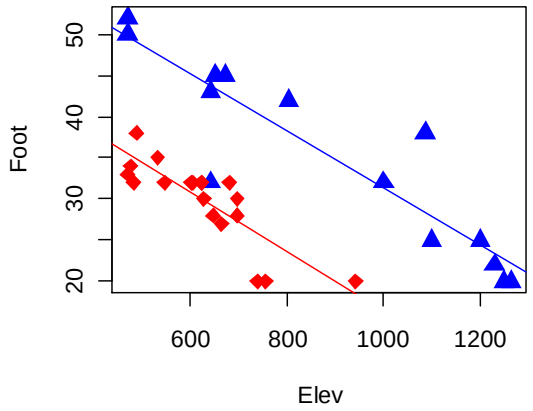
This is a method for factoring out the effects of conditions that are not part of the experimental design but which are there and can be measured.

T. Eden and R. A. Fisher. 1927. Studies in crop variation: IV. The experimental determination of the value of top dressings with cereals. The Journal of Agricultural Science 17: 548-562

Crested ibis nest site: footprint ~ landcover + elevation

Footprint	Elevation	Land cover	Nest site
22	1230	11	金家村
32	605	16	七氏山后
33	471	16	纸坊街5组
25	1200	11	3组1
32	602	16	7组
28	698	16	2组滚沟
50	471	11	3组2
52	471	11	蔡河4组
38	490	16	3组龙泉
28	648	16	牛河
20	942	16	代家河
20	1250	11	草坝4组
34	477	16	4组云阳
20	1264	11	华阳中学1号
32	681	16	4组黄沟
30	629	16	曹沟
32	483	16	5组麻洞
43	643	11	3组分会田
32	643	11	3组堰岔弯
30	698	16	后沟
28	698	16	汤帽
25	1100	11	草坝5组
35	533	16	党河电站
38	1087	11	高峰5组
45	674	11	7组袁沟
32	624	16	3组
20	757	16	沙溪沟
32	548	16	夏组
45	653	11	1组石洽
27	665	16	2组狗家沟
32	624	16	戴家沟
20	739	16	池塘岸
32	1001	11	8组
42	805	11	2组

Code	Land cover	类型
11	**Evergreen Needleleaf forest**	**常绿针叶林**
12	Deciduous Needleleaf forest	落叶针叶林
13	Evergreen Broadleaf forest	常绿阔叶林
14	Deciduous Broadleaf forest	落叶阔叶林
15	Mixed Froest	混交林
16	**Shrub**	**灌木**
21	Dense Grass	高覆盖度草地
22	Grass with Moderate Dense	中覆盖度草地
23	Sparse Grass	低覆盖度草地
31	Farmlamd	耕地
41	City and Urban Built-up	城市及建设用地
51	Harsh Desert	荒漠
52	Desert	沙漠
53	Bare Rock	裸露岩石
61	Wetland	湿地
62	Ice and Snow	冰川雪被
63	Waterbody	水体



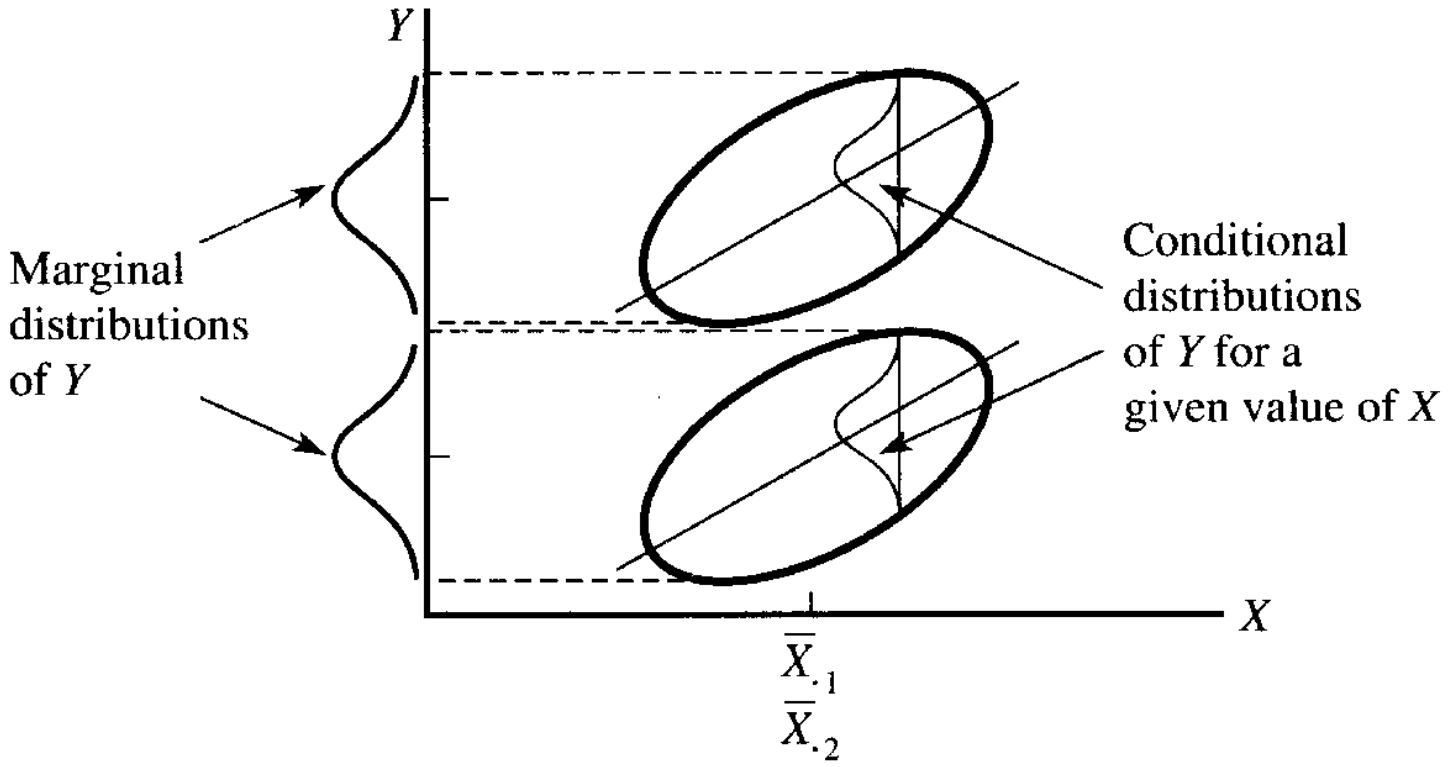
ANCOVA

- Analysis of Covariance
  - Combined use of ANOVA and Regression
    - Adjust for covariate by regressing covariate on the DV, then doing an ANOVA on the adjusted DV.

DV = IV x CV
DV: dependent variable (y, continuous)
IV: independent variable (x, categorical)
CV: covariate variable (x, continuous)

Two groups have same X , different Y

DV = IV x CV

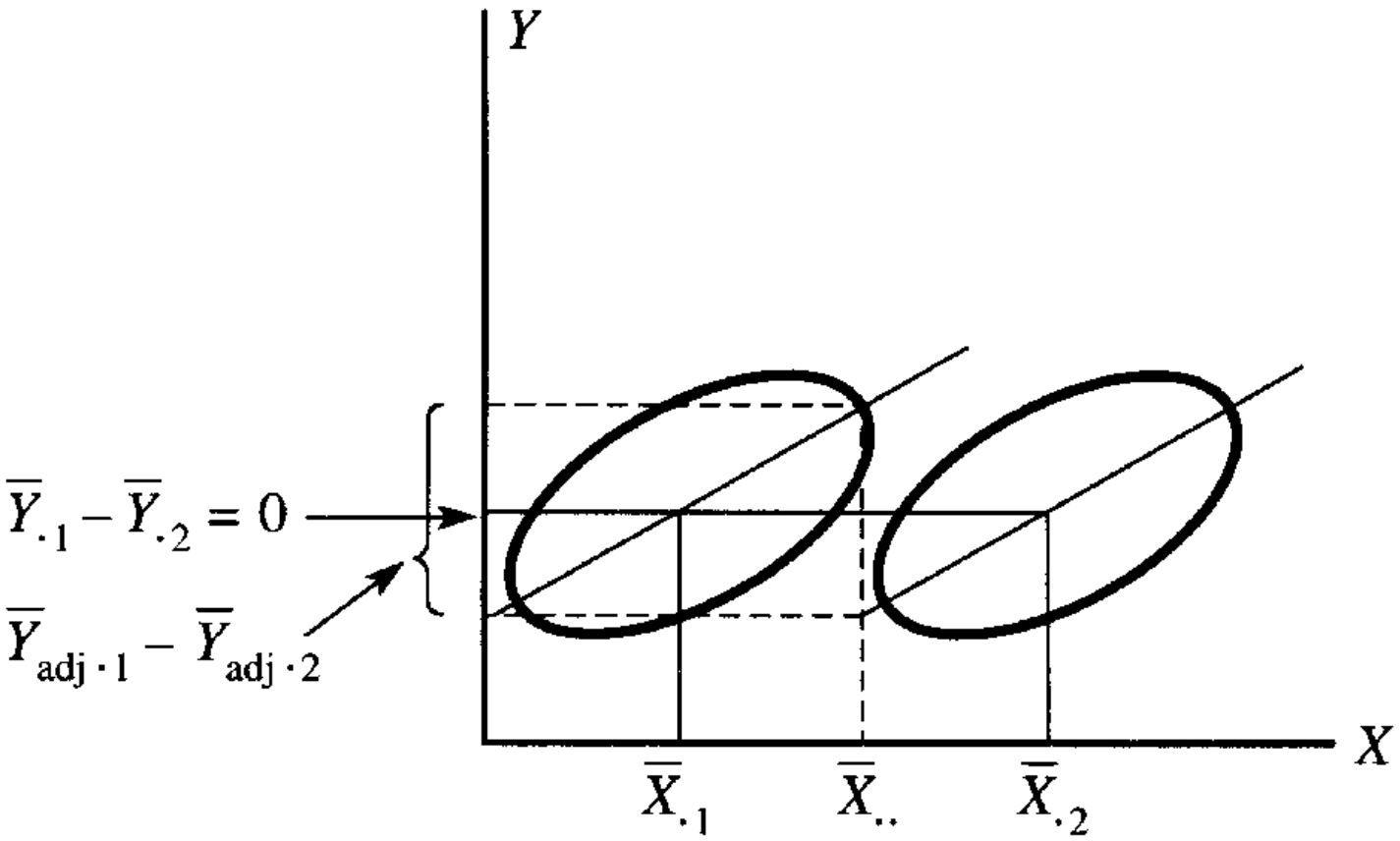


http://en.wikipedia.org/wiki/Marginal_distribution

In probability theory and statistics, the marginal distribution of a subset of a collection of random variables is the probability distribution of the variables contained in the subset. The term marginal variable is used to refer to those variables in the subset of variables being retained. These terms are dubbed "marginal" because they used to be found by summing values in a table along rows or columns, and writing the sum in the margins of the table. The distribution of the marginal variables (the marginal distribution) is obtained by marginalizing over the distribution of the variables being discarded, and the discarded variables are said to have been marginalized out.

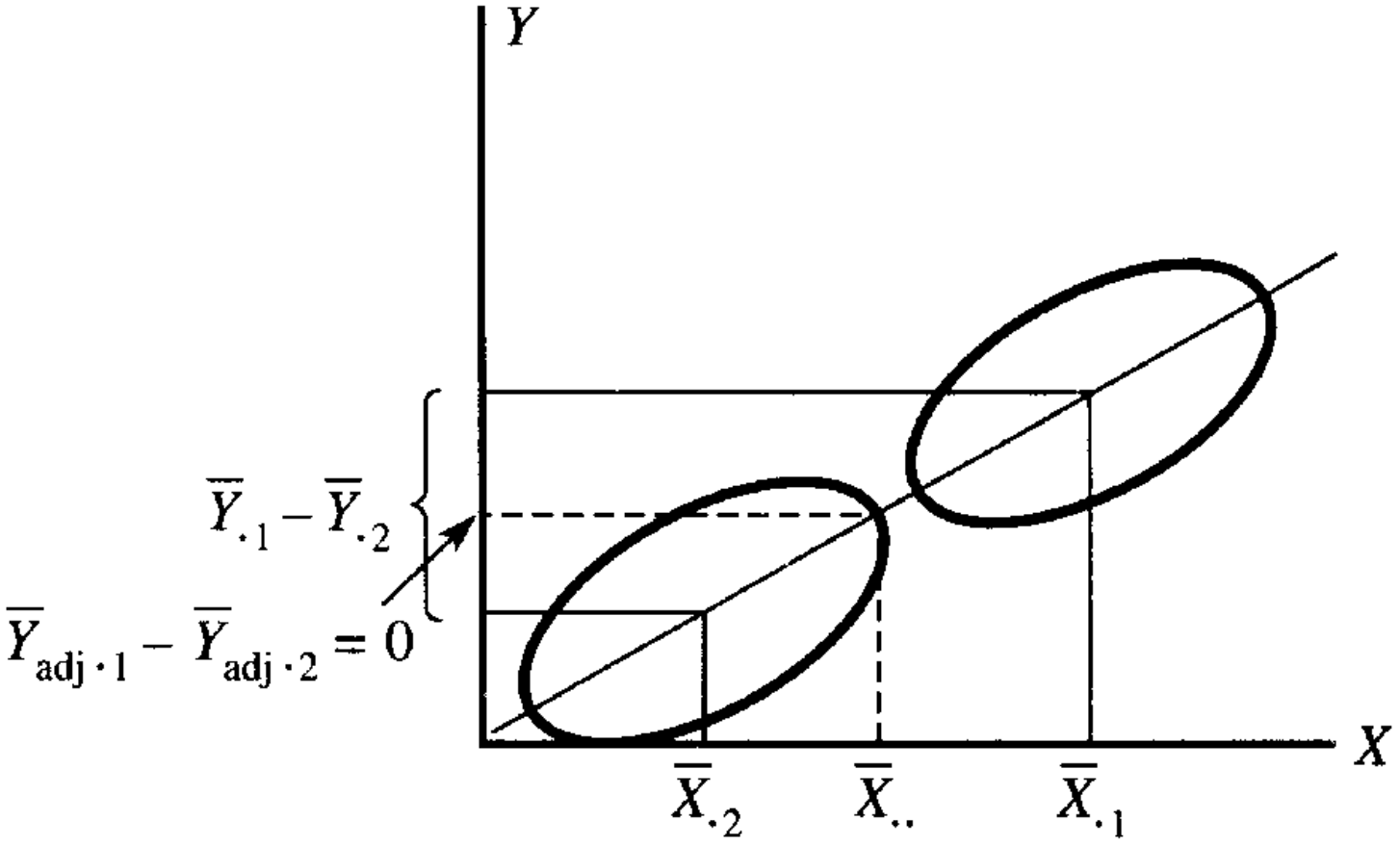
Two groups have different X with same Y, but the adjusted Y are different.

DV = IV x CV



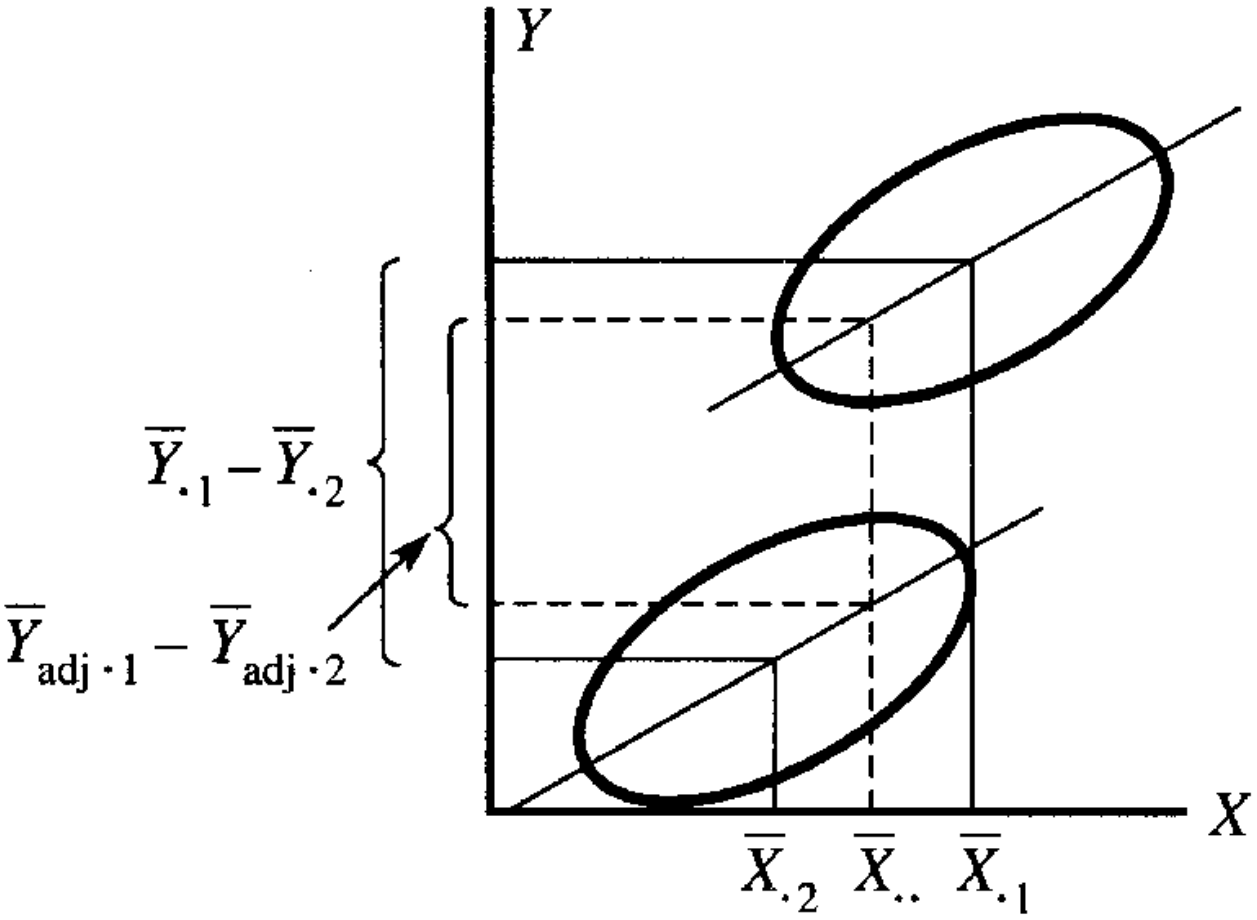
Two groups have different X , the adjusted Y are same

DV = IV x CV

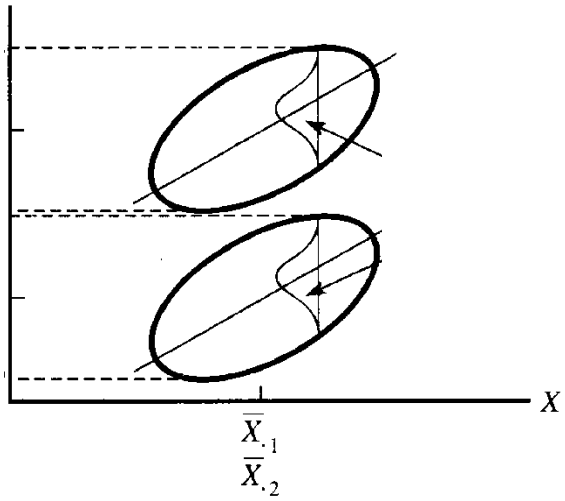


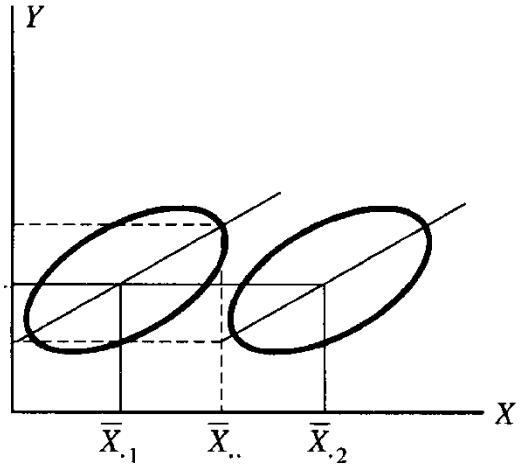
Two groups have different X , the adjusted Y are different

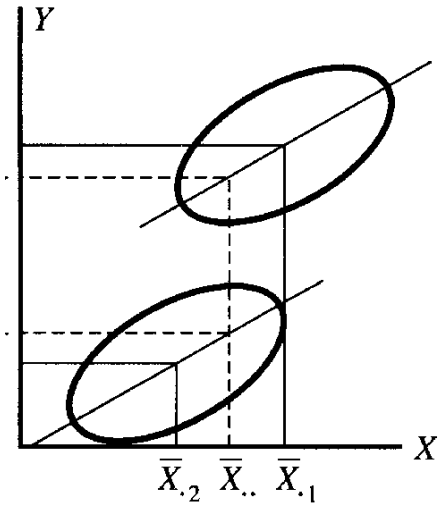
DV = IV x CV

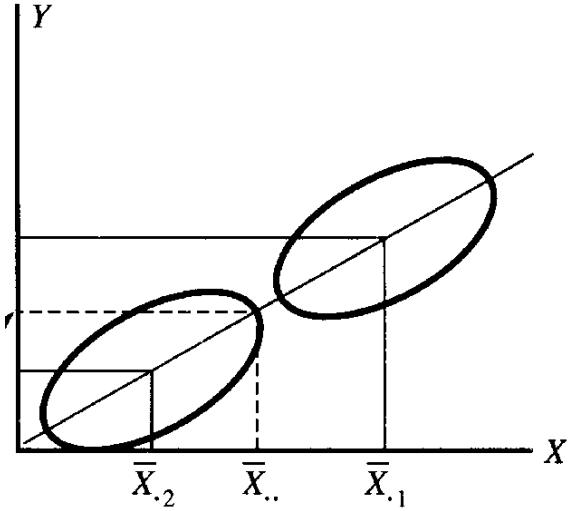


Real effect after adjustment





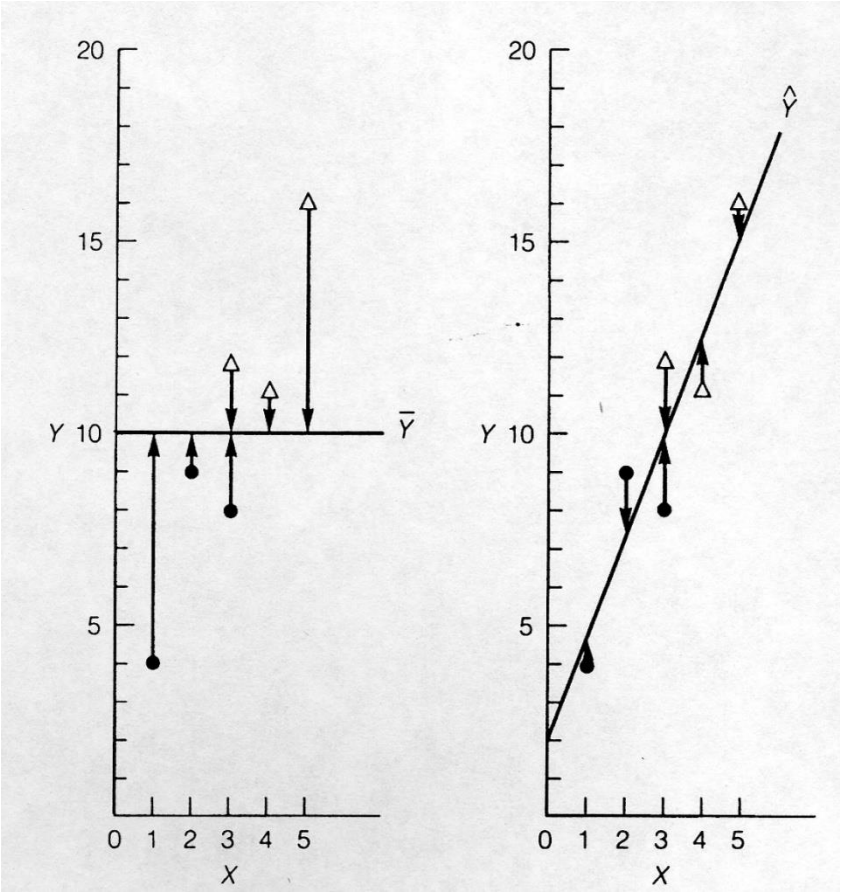




ANCOVA model

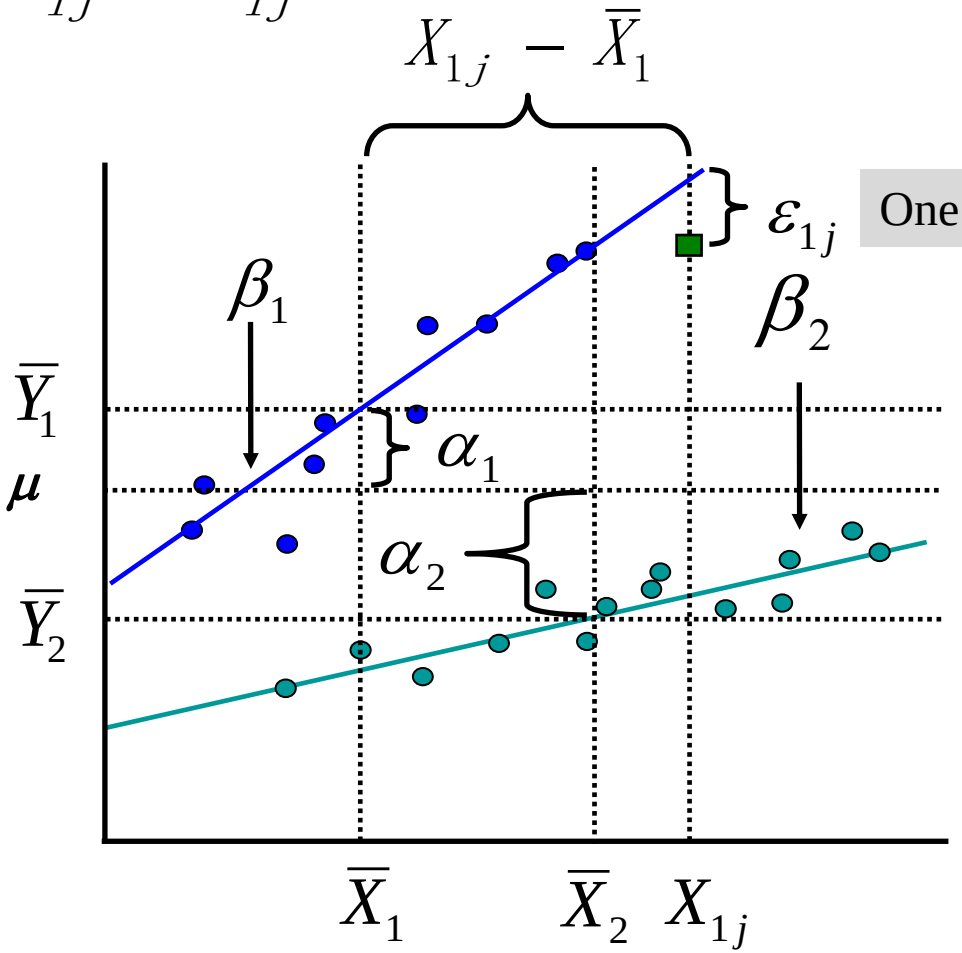
- Full model: $Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \varepsilon_{ij}$

- ANCOVA models have a major advantage over ANOVA
  - ANCOVA models have the capability of making a ***different prediction for each individual***, rather than having to make the same prediction for all individuals within a group
  - Predictions are a function of the score of the ***covariate*** X_{ij}



Model effect

$$Y_{ij} = \mu + \alpha_j + \beta_j X_{ij} + \varepsilon_{ij}$$



Assumptions

Normality of sampling distribution

- Normality on the DV at all of the levels of the IV(s) and the CV(s).
  - This cannot be shown unless you take multiple samples and form sampling distribution.

Homogeneity of Variance

- Equal variances on the DV at all of the levels of the IV(s) and the CV(s).

$$\sigma_1^2 \cong \sigma_2^2 \cong \ldots \sigma_p^2$$

  - This is most important after adjustments have been made, but if you have it before adjustment you are likely to have it afterwards.
  - If the assumption of homogeneity of variance fails, a more stringent alpha can be used (e.g. 0.01) or drop the variable from the analysis

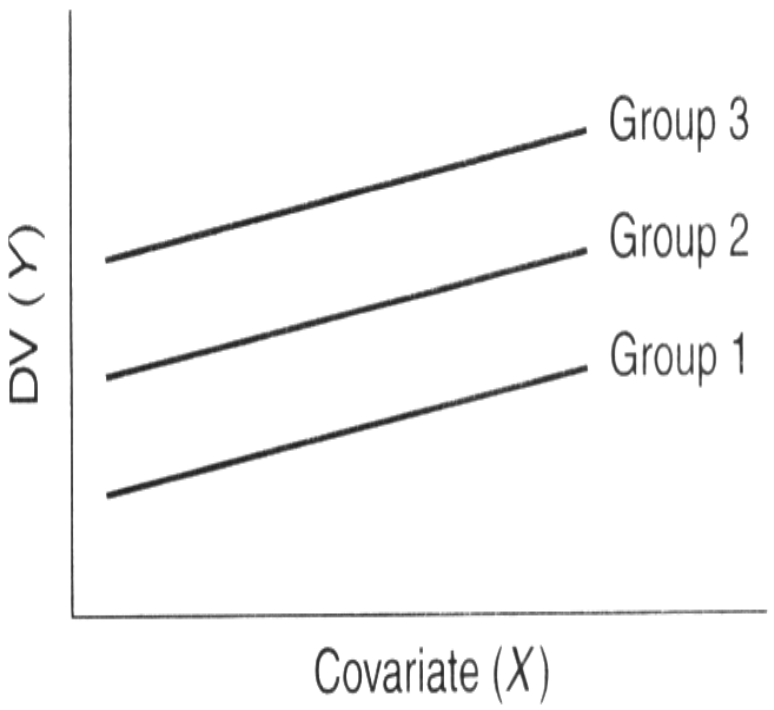
Linearity

- It is assumed that each DV has a linear relationship with the CV and other CVs.

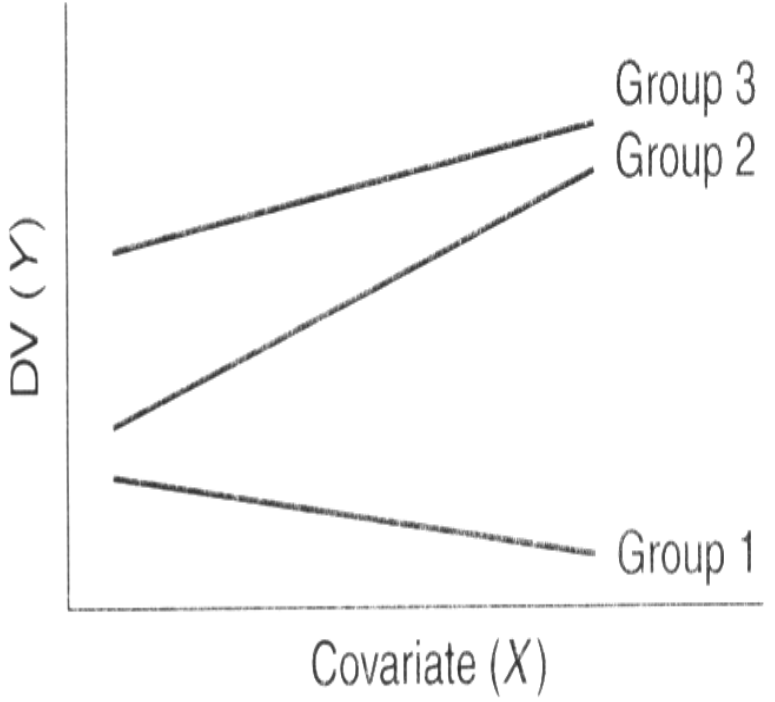
Homogeneity of regression

- For many cases, the slopes of the lines predicting the DV from the CV are same for each level of the IV.
  - In other words the regression coefficient (B) relating a CV to the DV should be the same for each group.
  - In still other words, this means no IV by DV interaction.

Homogeneity of Regression



(a) Homogeneity of regression (slopes)



(b) Heterogeneity of regression (slopes)

Dealing with heterogeneous within-group regression slopes

When slopes are clearly heterogeneous

- First, if the slopes themselves are of primary interest, you can contrast slopes across treatment combinations.
- Second, if the treatment (group) effects are the main interest, you can choose certain values of covariate and compare groups at those specific values.
  - e.g. using the mean of *X* or the value of *X* for which the distance between regression lines has the most precision.
- Third, use mixed effect models to quantify the effects of IV and CV.

Reliability of Covariates

- It is assumed that each CV is measured without error (this is unrealistic).

- So it is recommended that CVs only be used when they meet a reliability of .8 or more.

Outlier

- No outliers
  - Test for univariate outliers on the DV and all of the CVs individually
  - Test for multivariate outliers in the combined DV and CVs space.

Outlier

- A data point that is distinctly separate from the rest of the data. One definition of outlier is any data point more than 1.5 interquartile ranges (IQRs) below the first quartile or above the third quartile.
- Note: The IQR definition given here is widely used but is not the last word in determining whether a given number is an outlier.

http://www.mathwords.com/o/outlier.htm

Outlier

Example: For the data 2, 5, 6, 9, 12

minimum = 2
first quartile = 3.5
median = 6
third quartile = 10.5
maximum = 12

IQR = 10.5 – 3.5 = 7, so 1.5·IQR = 10.5.

To determine if there are outliers we must consider the numbers that are 1.5·IQR or 10.5 beyond the quartiles.

first quartile – 1.5·IQR = 3.5 – 10.5 = –7
third quartile + 1.5·IQR = 10.5 + 10.5 = 21

Since none of the data are outside the interval from –7 to 21, there are no outliers.

http://www.mathwords.com/o/outlier.htm

No Multicollinearity

- If a CV is highly related to another CV (e.g. at a correlation of .5 or more) then it should not be used to adjust the DV over the other CV.

- One or the other should be removed since they are statistically redundant.

Unequal sample sizes, missing data and number of cases

- The problem here is that with unequal samples it is unclear how to calculate the marginal mean.

- Another problem is that the variances then start to overlap one another forcing the within plus between variances to be larger than the total variance.

Type I and III sum of squares: examples

1, 1 3, 3 Mean = 2

1, 1
1, 1
1, 1 3, 3 Mean = 1.5

Type I sum of squares use 1.5 as the global mean;
Type III sum of square use 2 as the global mean.

Unequal sample sizes, missing data and number of cases

```
# Tensile strength in paper manufacturing
Y <- c(30,35,37,36,34,41,38,42,29,26,33,36,
      28,32,40,41,31,36,42,40,31,30,32,40,
      31,37,41,40,35,40,39,44,32,34,39,45)

block <- gl(3, 12, 36) # Three blocks
A     <- gl(3, 4, 36) # Three pulp preparation methods
B     <- gl(4, 1, 36) # Four different temperatures
Dat   <- data.frame(Y, block, A, B)

Dat = Dat[-c(1:3), ] # make data unbalanced


summary(aov(Y ~ A*B, data = Dat)) # type I sum of square
Anova(mod <- lm(Y ~ A*B, data = Dat),  type = "II") # library(car)
Anova(mod <- lm(Y ~ A*B, data = Dat),  type = "III")
```

	Y	block	A	B
[1,]	30	1	1	1
[2,]	35	1	1	2
[3,]	37	1	1	3
[4,]	36	1	1	4
[5,]	34	1	2	1
[6,]	41	1	2	2
[7,]	38	1	2	3
[8,]	42	1	2	4
[9,]	29	1	3	1
[10,]	26	1	3	2
[11,]	33	1	3	3
[12,]	36	1	3	4
[13,]	28	2	1	1
[14,]	32	2	1	2
[15,]	40	2	1	3
[16,]	41	2	1	4
[17,]	31	2	2	1
[18,]	36	2	2	2
[19,]	42	2	2	3
[20,]	40	2	2	4
[21,]	31	2	3	1
[22,]	30	2	3	2
[23,]	32	2	3	3
[24,]	40	2	3	4
[25,]	31	3	1	1
[26,]	37	3	1	2
[27,]	41	3	1	3
[28,]	40	3	1	4
[29,]	35	3	2	1
[30,]	40	3	2	2
[31,]	39	3	2	3
[32,]	44	3	2	4
[33,]	32	3	3	1
[34,]	34	3	3	2
[35,]	39	3	3	3
[36,]	45	3	3	4

Unequal sample sizes, missing data and number of cases

ANOVA Table (Type I tests)					
	Df	**Sum Sq**	**Mean Sq**	**F value**	**P**
A	2	126	63.02	7.484	0.00352
B	3	398.9	132.96	15.79	1.33E-05
A:B	6	81.8	13.63	1.618	0.19147
Residuals	21	176.8	8.42		
ANOVA Table (Type II tests)					
	Df	**Sum Sq**	**Mean Sq**	**F value**	**P**
A	2	127.5		7.5709	0.003343
B	3	398.88		15.7898	1.33E-05
A:B	6	81.76		1.6182	0.191471
Residuals	21	176.83			
ANOVA Table (Type III tests)					
	Df	**Sum Sq**	**Mean Sq**	**F value**	**P**
(Intercept)	1	1740.5		206.6946	2.43E-12
A	2	20.04		1.19	0.323909
B	3	156.06		6.1775	0.003539
A:B	6	81.76		1.6182	0.191471
Residuals	21	176.83			

Calculating type I, II, and III sum of squares

Types of sum of squares	Variables and terms	Sum of squares
Type I SS	X1	SS(X1)
	X2	SS(X2 | X1)
	X1:X2	SS(X1X2 | X2, X1)
Type II SS	X1	SS(X1 | X2)
	X2	SS(X2 | X1)
	X1:X2	SS(X1X2 | X2, X1)
Type III SS	X1	SS(X1 | X2, X1X2)
	X2	SS(X2 | X1, X1X2)
	X1:X2	SS(X1X2 | X2, X1)

SS(X1X2 | X1, X2) = SS(X1, X2, X1X2) – SS(X1, X2)
SS(X1 | X2, X1X2) = SS(X1, X2, X1X2) – SS(X2, X1X2)
SS(X2 | X1, X1X2) = SS(X1, X2, X1X2) – SS(X1, X1X2)
SS(X1 | X2) = SS(X1, X2) – SS(X2)
SS(X2 | X1) = SS(X1, X2) – SS(X1)

Calculating type I, II, and III sum of squares

```
data(mtcars)
mtcars$cyl = as.factor(mtcars$cyl)
mtcars$am = as.factor(mtcars$am)
table(mtcars$cyl)
table(mtcars$am)

# SS of mpg
SSE = sum(residuals(lm(mpg ~ 1, data=mtcars))^2) # 1126
# SSE of mpg after cyl being explained
SSE.cyl   = sum(residuals(lm(mpg ~ cyl, data=mtcars))^2) # 301
# SSE of mpg after cyl and am being explained
SSE.cyl.am = sum(residuals(lm(mpg ~ cyl+am, data=mtcars))^2) # 264
# SSE of mpg after cyl, am, and cyl:am being explained
SSE.cyl.am.cyl_am= sum(residuals(lm(mpg ~ cyl*am, data=mtcars))^2) # 239

## Type I SS
SS.cyl = SSE - SSE.cyl      # SS of cyl, 825
SS.am = SSE.cyl - SSE.cyl.am  # SS of am, 37
SS.cyl_am = SSE.cyl.am - SSE.cyl.am.cyl_am # SS of cyl:am, 25
```

Unequal sample sizes, missing data and number of cases

- Do not use type 1 sums of squares
- Type 1 sums of squares assumes that the difference in number of subjects is meaningful and gives more weight to the values from larger groups
- Order of variables (X1, X2) matters.

Unequal sample sizes, missing data and number of cases

- Use the type 3 (III) sums of square
- The type 3 sums of square assumes that the data was supposed to be complete, and the difference in the number of subjects is not meaningful
  - Acts like standard multiple regression. Each main effect and interaction is assessed after all other main effects, interactions and covariates are controlled
  - Treats all groups the same – small group is weighted equally as a large group (sometimes called the unweighted approach)
  - Are preferable in most cases since they correspond to the variation attributable to an effect after correcting for any other effects in the model. They are unaffected by the frequency of observations.
  - Order of variables (X1, X2) does not matter.

Unequal sample sizes, missing data and number of cases

- Number of cases required depends on the number needed to reach appropriate level of power.
- Unbalanced experimental design needs more cases.

Advantages of ANCOVA

- Adjusts for pre-treatment differences between groups.

  If pre-treatment differences exist because groups were not randomly formed, then ANCOVA will eliminate the bias that may exist with non-random assignment.

- More Power – due to decreased variance that must be explained by the IV (smaller error term in the F test).

  Covariate “accounts for” some of the variance in the DV variance.

Equations (model: Y = X x Z)

- Just like ANOVA the total variance can be separated into within and between groups variance:

$$\sum_i \sum_j (Y_{ij} - GM_{(y)})^2 = n\sum_j \left(\bar{Y}_j - GM_{(y)}\right)^2 + \sum_i \sum_j (Y_{ij} - \bar{Y}_j)^2$$

$$SS_{Total(y)} = SS_{bg(y)} + SS_{wg(y)}$$

GM: grand mean

Equations (model: Y = X × Z)

- But in ANCOVA you also have a partitioning of the variance in each CV:

$$\sum_i \sum_j (Z_{ij} - GM_{(z)})^2 = n \sum_j \left(\bar{Z}_j - GM_{(z)} \right)^2 + \sum_i \sum_j (Z_{ij} - \bar{Z}_j)^2$$

$$SS_{Total(z)} = SS_{bg(z)} + SS_{wg(z)}$$

Equations (model: Y = X × Z)

- And also a partitioning of the covariation between them (DV and CV):

$$SP_{Total} = SP_{bg} + SP_{wg}$$

Equations (model: Y = X × Z)

- This covariation is used to adjust the between and within groups sums of squares:

$$SS'_{bg(y)} = SS_{bg(y)} - \left[\frac{\left(SP_{bg} + SP_{wg} \right)^2}{SS_{bg(z)} + SS_{wg(z)}} - \frac{\left(SP_{wg} \right)^2}{SS_{wg(z)}} \right]$$

$$SS'_{wg(y)} = SS_{wg(y)} - \frac{\left(SP_{wg} \right)^2}{SS_{wg(z)}}$$

Equations (model: Y = X × Z)

The adjustment made to the between group scores in last slide can also be conceptualized as:

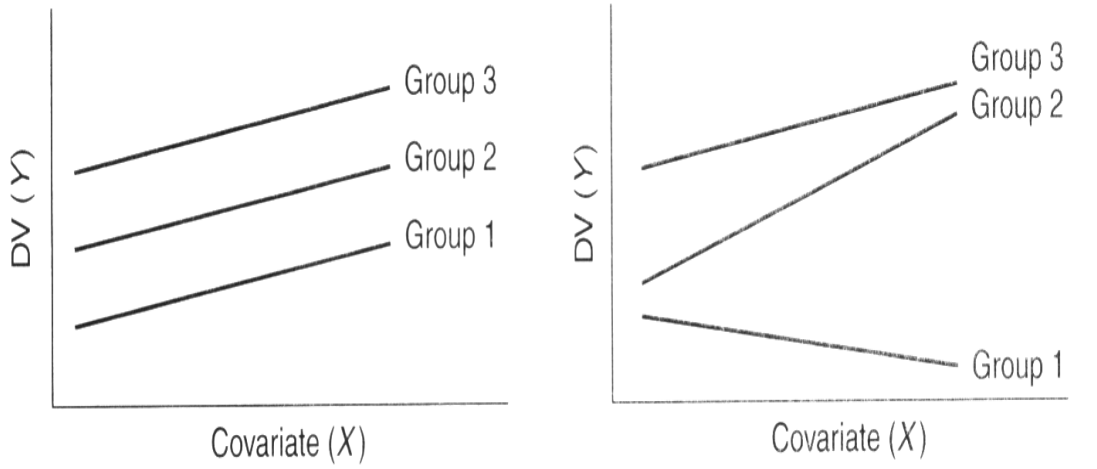
the adjustment is obtained by

- taking each individuals deviation around Y grand mean (before adjustment)
- subtracting from it each person's deviation around the Z grand mean, weighted by the relationship between the two variables

$$(Y - \bar{Y}_j) = (Y - GM_{(y)}) - \beta_{y \cdot z}(Z - GM_{(z)})$$

Degree of freedom

- Each regression coefficient (slope) we need to estimate (one for every CV) eats up a degree of freedom (if slopes are different).
- This accounts for the smaller error degrees of freedom when compared to regular ANOVA.



(a) Homogeneity of regression (slopes) (b) Heterogeneity of regression (slopes)

General issues

The basic ANCOVA tests null hypotheses about adjusted factor effects, where the linear relationship between the covariate and the response variable (*Y*) is taken into account.

These means at all levels of factors are adjusted to the overall mean value for the covariate by the relationship between *Y* and the covariate.

General issues

Homogeneity of within-group regression slopes is tested by including factor by covariate interaction terms in a preliminary model.

In complex models with many categorical explanatory variables, homogeneity of slopes can be checked by combining all factors by the covariate terms into a single interaction term.

Alternatively, homogeneity of slopes may be better tested separately for each component of the analysis.

Alternatives to ANCOVA

- When CV and DV have a non-linear relationship
  - Use CV to group similar observations together into blocks. Each block is then used as levels of a BG IV that is crossed with the other BG IV that you are interested in.
  - Blocking may be the best alternative, because it doesn't have the special assumptions of ANCOVA, and it can capture non-linear relationships between CV and DV where ANCOVA only deals with linear relationships.

Example (crested ibis nest site): footprint ~ landcover + elevation

Footprint	Elevation	Land cover	Nest site
22	1230	11	金家村
32	605	16	七氏山后
33	471	16	纸坊街5组
25	1200	11	3组1
32	602	16	7组
28	698	16	2组滚沟
50	471	11	3组2
52	471	11	蔡河4组
38	490	16	3组龙泉
28	648	16	牛河
20	942	16	代家河
20	1250	11	草坝4组
34	477	16	4组云阳
20	1264	11	华阳中学1号
32	681	16	4组黄沟
30	629	16	曹沟
32	483	16	5组麻洞
43	643	11	3组分会田
32	643	11	3组堰岔弯
30	698	16	后沟
28	698	16	汤帽
25	1100	11	草坝5组
35	533	16	党河电站
38	1087	11	高峰5组
45	674	11	7组袁沟
32	624	16	3组
20	757	16	沙溪沟
32	548	16	夏组
45	653	11	1组石洽
27	665	16	2组狗家沟
32	624	16	戴家沟
20	739	16	池塘岸
32	1001	11	8组
42	805	11	2组

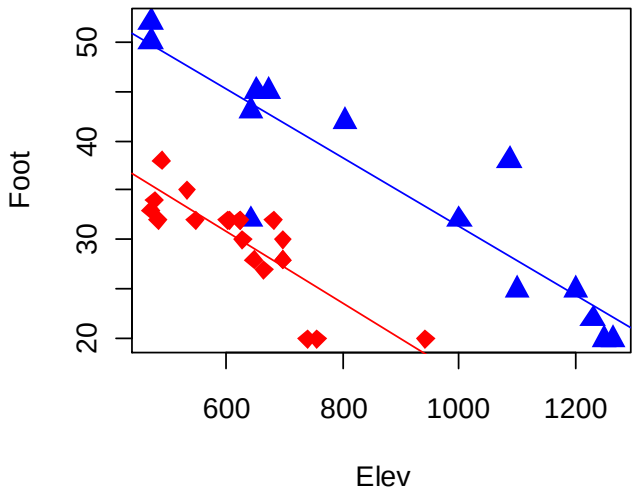
Code	Land cover	类型
11	Evergreen Needleleaf forest	常绿针叶林
12	Deciduous Needleleaf forest	落叶针叶林
13	Evergreen Broadleaf forest	常绿阔叶林
14	Deciduous Broadleaf forest	落叶阔叶林
15	Mixed Froest	混交林
16	Shrub	灌木
21	Dense Grass	高覆盖度草地
22	Grass with Moderate Dense	中覆盖度草地
23	Sparse Grass	低覆盖度草地
31	Farmlamd	耕地
41	City and Urban Built-up	城市及建设用地
51	Harsh Desert	荒漠
52	Desert	沙漠
53	Bare Rock	裸露岩石
61	Wetland	湿地
62	Ice and Snow	冰川雪被
63	Waterbody	水体

R code – plot data

```
#ANCOVA
#Human footprint index, elevation, and landcover
nrow(ibis) #34 nests
ibis$Landcover[ibis$Landcover == 11] <- 1 #Forest
ibis$Landcover[ibis$Landcover == 16] <- 2 #Shrub
ibis$Landcover <- as.factor(ibis$Landcover)

Elev = ibis$Elev; Foot = ibis$Foot; Landcover = ibis$Landcover
plot(Elev, Foot, pch=16+as.numeric(Landcover), col=c('blue',
     'red')[as.numeric(Landcover)],cex=1.5)
abline(lm(Foot[Landcover==1]~Elev[Landcover==1]),lty=1, col='blue')
abline(lm(Foot[Landcover==2]~Elev[Landcover==2]),lty=1, col='red')
```

```
lines <-
"Foot Elev Landcover Nestsite
22 1230 11 金家村
32 605 16 七氏山后
33 471 16 纸坊街5组
25 1200 11 3组1
32 602 16 7组
28 698 16 2组滚沟
50 471 11 3组2
52 471 11 板河4组
38 490 16 3组龙泉
28 648 16 牛河
20 942 16 代家河
20 1250 11 草坝4组
34 477 16 4组云阳
20 1264 11 华阳中学1号
32 681 16 4组黄沟
30 629 16 曹沟
32 483 16 5组麻洞
43 643 11 3组分会田
32 643 11 3组堰岔弯
30 698 16 后沟
28 698 16 汤帽
25 1100 11 草坝5组
35 533 16 党河电站
38 1087 11 高峰5组
45 674 11 7组袁沟
32 624 16 3组
20 757 16 沙溪沟
32 548 16 夏组
45 653 11 1组石沿
27 665 16 2组狗家沟
32 624 16 戴家沟
20 739 16 池塘岸
32 1001 11 8组
42 805 11 2组"
ibis <- read.table(con
    <- textConnection(lines),
    header=TRUE)
close(con)
```

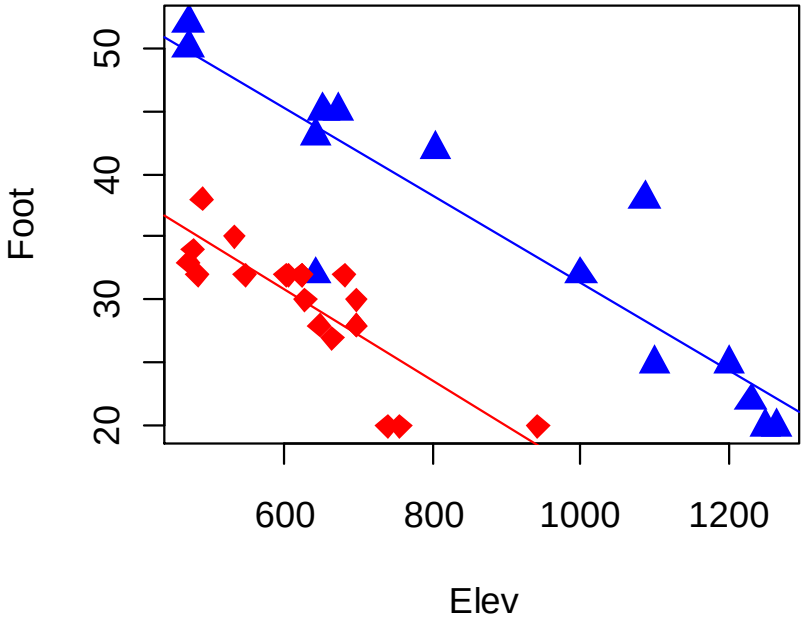
Foot	Elev	Landcover	Nestsite
22	1230	11	金家村
32	605	16	七氏山后



Compare means

```
options(digits=3)
tapply(Foot, Landcover, mean)
t.test(Foot ~ Landcover)
```

```
> tapply(Foot, Landcover, mean)
  1   2
35.1 29.8

> t.test(Foot ~ Landcover)

        Welch Two Sample t-test

data:  Foot by Landcover
t = 1.65, df = 16.5, p-value = 0.118
alternative hypothesis:
true difference in means is not equal to 0
95 percent confidence interval:
 -1.5 12.1
sample estimates:
mean in group 1 mean in group 2
           35.1            29.8
```



summary(ancova)

```
anova1 <- lm(Foot~Landcover)
summary(anova1)
```

```
Call:
lm(formula = Foot ~ Landcover)

Residuals:
   Min     1Q Median     3Q    Max
-15.07  -3.07   2.25   4.00  16.93

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)    35.07       2.18   16.07    <2e-16
Landcover2     -5.32       2.85   -1.87     0.071 .
---

Residual standard error: 8.17 on 32 degrees of freedom
Multiple R-squared: 0.0985,     Adjusted R-squared: 0.0703
F-statistic:  3.5 on 1 and 32 DF,  p-value: 0.0707
```

```
ancova <- lm(Foot~Landcover*Elev)
summary(ancova)
```

```
Call:
lm(formula = Foot ~ Landcover * Elev)

Residuals:
    Min      1Q  Median      3Q     Max
-11.702  -1.475   0.634   1.939   9.670

Coefficients:
                Estimate   Std. Error  t value    Pr(>|t|)
(Intercept)       65.96    3.24635     20.32      < 2e-16
Landcover2       -13.41    5.76795     -2.33       0.027
Elev              -0.034    0.00346   -10.00       4.6e-11
Landcover2:Elev  -0.00153  0.00821     -0.19       0.853
---

Residual standard error: 3.73 on 30 degrees of freedom
Multiple R-squared: 0.824,      Adjusted R-squared: 0.806
F-statistic: 46.8 on 3 and 30 DF,  p-value: 1.99e-11
```

ANOVA table

anova(ancova)

Analysis of Variance Table

Response: Foot

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Landcover	1	233	233	16.77	0.00029 ***
Elev	1	1717	1717	123.52	3.7e-12 ***
Landcover:Elev	1	0	0	0.03	0.85316
Residuals	30	417	14		

Update model

```
ancova2 = update(ancova,  ~. -Landcover:Elev)
anova(ancova, ancova2)
```

```
Analysis of Variance Table

Model 1: Foot ~ Landcover * Elev
Model 2: Foot ~ Landcover + Elev

  Res.Df  RSS  Df  Sum of Sq    F    Pr(>F)
1     30  417
2     31  418  -1    -0.485  0.03    0.85
```

Compare with ANOVA

```
ancova3 = update(ancova2, ~. -Elev)
anova(ancova2, ancova3)
```

```
Analysis of Variance Table

Model 1: Foot ~ Landcover + Elev
Model 2: Foot ~ Landcover
  Res.Df    RSS Df      Sum of Sq   F       Pr(>F)
1     31       418
2     32      2135 -1     -1717       127     1.6e-12 ***
```

Model selection

step(ancova)

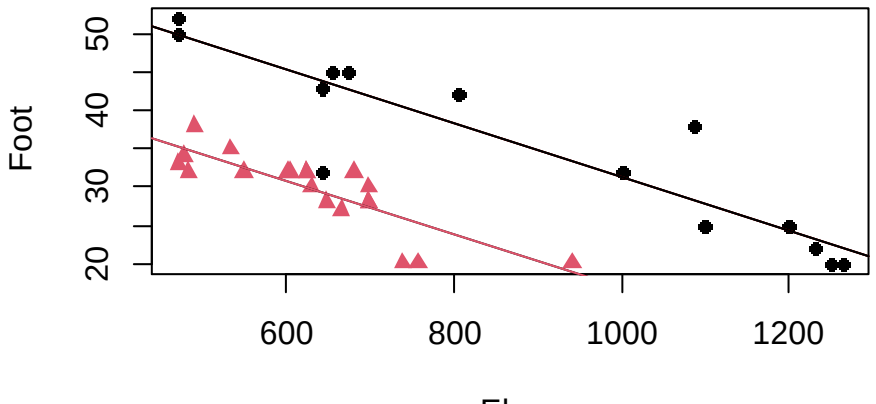
```
Start:  AIC=93.2
Foot ~ Landcover * Elev

                 Df  Sum of Sq  RSS  AIC
- Landcover:Elev  1     0.485  418  91.3
<none>                         417  93.2

Step:  AIC=91.3
Foot ~ Landcover + Elev

            Df  Sum of Sq   RSS   AIC
<none>                     418   91.3
- Landcover  1      1229  1646 135.9
- Elev       1      1717  2135 144.8

Call:
lm(formula = Foot ~ Landcover + Elev)

Coefficients:
(Intercept)   Landcover2         Elev
    66.2053     -14.4522      -0.0349
```

Model coefficients

```
ancova <- lm(Foot~Landcover+Elev)
summary(ancova)
```

```
Coefficients:
             Estimate   Std. Error  t value    Pr(>|t|)
(Intercept)  66.20532   2.92669     22.621     < 2e-16 ***
Landcover2  -14.45224   1.51312     -9.551     9.48e-11 ***
Elev         -0.03489   0.00309     -11.291    1.63e-12 ***
```

```
plot(Foot~Elev, pch=as.numeric(Landcover)+15, col=as.numeric(Landcover))
abline (66.20532, -0.03489, col = 1 ) #
abline (66.20532-14.45224,  -0.03489, col=2) #
```



Akaike information criterion (AIC)

The Akaike information criterion is a measure of the relative goodness of fit of a statistical model. It was developed by Hirotsugu Akaike in 1974.

AIC provides a means for comparison among models, a tool for model selection.

$$AIC = 2k - 2\ln(L)$$

where k is the number of parameters in the statistical model, and L is the maximized value of the likelihood function for the estimated model

AICc is AIC with a correction for finite sample sizes

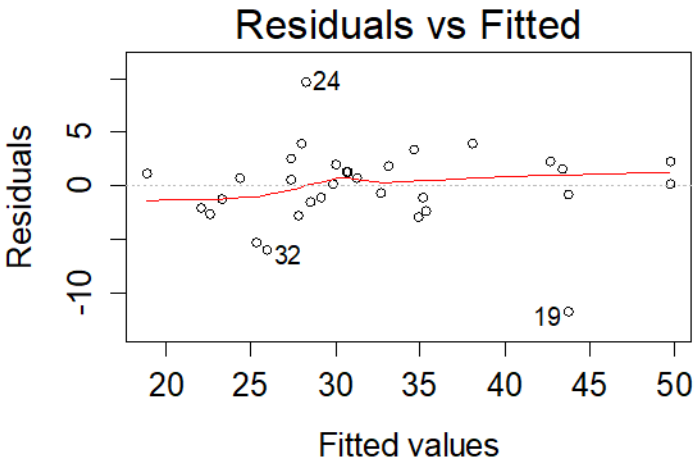
$$AIC_C = AIC + \frac{2k(k+1)}{n-k-1}$$

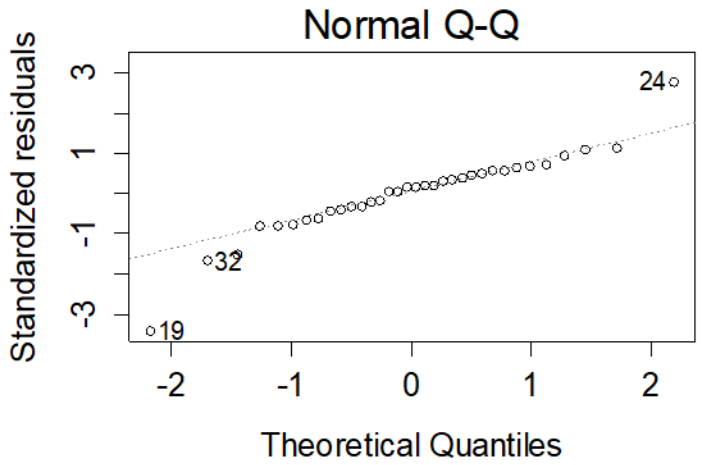
where k denotes the number of model parameters. Thus, AICc is AIC with a greater penalty for extra parameters.

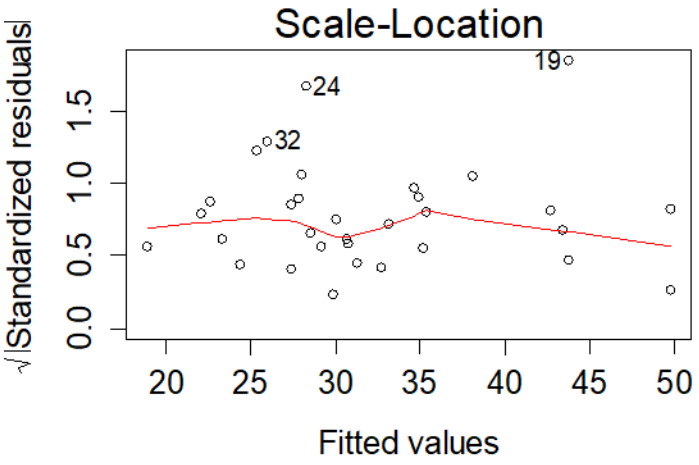
Akaike, Hirotugu (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6): 716–723.

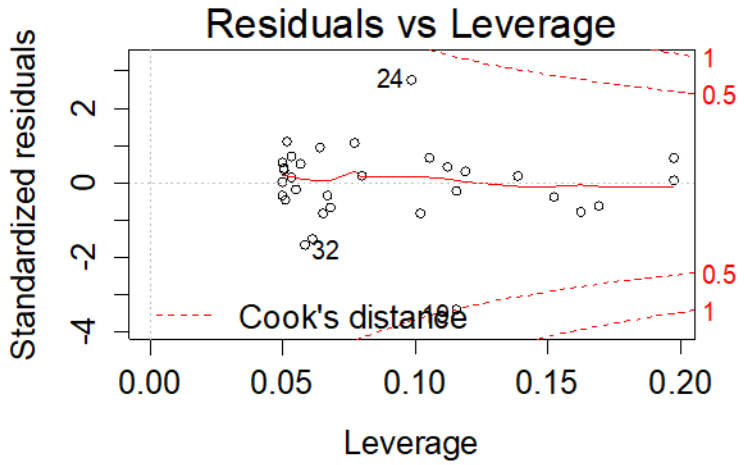
Model check

```
ancova.final <- step(ancova)
plot(ancova.final)
```









1 1 3 10 122
2 2 24 34 129
3 3 10 21 114
4 1 5 10 121
5 2 18 27 114
6 3 3 18 114
7 1 6 14 101
8 2 11 20 116
9 3 10 20 110
10 1 11 29 131
11 2 10 13 126
12 3 3 9 94
13 1 11 17 129
14 2 11 19 110
15 3 6 13 102
16 1 13 21 115
17 2 2 28 138
18 3 9 24 128
19 1 7 5 122
20 2 10 13 119
21 3 13 19 111
22 1 12 17 112
23 2 14 21 123
24 3 7 25 119
25 1 13 17 123
26 2 11 14 115
27 3 10 24 120
28 1 8 22 119
29 2 12 17 116
30 3 9 21 112
31 1 9 22 122
32 2 14 16 125
33 3 7 21 105
34 1 10 18 111
35 2 7 10 122
36 3 4 17 120
37 1 6 11 117
38 2 8 18 120
39 3 7 24 120
40 1 13 20 112
41 2 10 13 111
42 3 12 25 118
43 1 7 8 122
44 2 11 17 127
45 3 6 23 110
46 1 11 20 124
47 2 12 13 122
48 3 7 22 127
49 1 5 15 118
50 2 6 13 127
51 1 9 25 113
52 2 3 13 115
53 1 8 25 126
54 2 4 13 112
55 1 2 14 132
56 1 11 17 93

post.score ~ class.type * pre.score * IQ

/* Data for the ANCOVA example (the Trigonometry scores) */

ID	Class type	pre score	post score	IQ
1	1	3	10	122
2	2	24	34	129
3	3	10	21	114
4	1	5	10	121
5	2	18	27	114
6	3	3	18	114

lines <-
"obs class pre post IQ
1 1 3 10 122
2 2 24 34 129
3 3 10 21 114
4 1 5 10 121
5 2 18 27 114
6 3 3 18 114
7 1 6 14 101
8 2 11 20 116
9 3 10 20 110
10 1 11 29 131
11 2 10 13 126
12 3 3 9 94
13 1 11 17 129
14 2 11 19 110
15 3 6 13 102
16 1 13 21 115
17 2 2 28 138
18 3 9 24 128
19 1 7 5 122
20 2 10 13 119
21 3 13 19 111
22 1 12 17 112
23 2 14 21 123
24 3 7 25 119
25 1 13 17 123
26 2 11 14 115
27 3 10 24 120
28 1 8 22 119
29 2 12 17 116
30 3 9 21 112
31 1 9 22 122
32 2 14 16 125
33 3 7 21 105
34 1 10 18 111
35 2 7 10 122
36 3 4 17 120
37 1 6 11 117
38 2 8 18 120
39 3 7 24 120
40 1 13 20 112
41 2 10 13 111
42 3 12 25 118
43 1 7 8 122
44 2 11 17 127
45 3 6 23 110
46 1 11 20 124
47 2 12 13 122
48 3 7 22 127
49 1 5 15 118
50 2 6 13 127
51 1 9 25 113
52 2 3 13 115
53 1 8 25 126
54 2 4 13 112
55 1 2 14 132
56 1 11 17 93"
scores <- read.table(con <- textConnection(lines), header=TRUE); close(con)

ancova = lm(post.score ~ class.type * pre.score * IQ)

Results

```
scores$class = factor(scores$class)
ancova = lm(post ~ class * pre * IQ,
            data = scores)
summary(ancova)
```

Call:
lm(formula = post ~ class * pre * IQ, data = scores)

Residuals:

Min	1Q	Median	3Q	Max
-10.771	-2.900	-0.288	3.055	8.153

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.12485	43.56642	-0.03	0.98
class2	-63.85851	54.87780	-1.16	0.25
class3	-29.91929	60.51750	-0.49	0.62
pre	0.50020	4.48630	0.11	0.91
IQ	0.07686	0.35743	0.22	0.83
class2:pre	4.60531	5.57265	0.83	0.41
class3:pre	1.75546	7.85225	0.22	0.82
class2:IQ	0.53693	0.44687	1.20	0.24
class3:IQ	0.34119	0.52100	0.65	0.52
pre:IQ	0.00471	0.03702	0.13	0.90
class2:pre:IQ	-0.04029	0.04551	-0.89	0.38
class3:pre:IQ	-0.02008	0.06813	-0.29	0.77

Residual standard error: 4.66 on 44 degrees of freedom
Multiple R-squared: 0.484, Adjusted R-squared: 0.355
F-statistic: 3.76 on 11 and 44 DF, p-value: 0.000793

Model selection

```
ancova2 = update(ancova, ~. -class : pre : IQ)
summary(ancova2)
ancova3 = update(ancova2, ~. -class : pre)
summary(ancova3)
ancova4 = update(ancova3, ~. -class : IQ)
summary(ancova4)
ancova5 = update(ancova4, ~. -pre : IQ)
summary(ancova5)
```

```
Call:
lm(formula = post ~ class + pre + IQ, data = scores)

Coefficients:
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept) -14.8759     8.8927   -1.67     0.1005
class2       -1.4026     1.4889   -0.94     0.3506
class3        4.9870     1.5609    3.20     0.0024 **
pre           0.7802     0.1596    4.89     1e-05 ***
IQ            0.2129     0.0736    2.89     0.0056 **
```

Model selection

step(ancova)

```
Start:  AIC=183
post ~ class * pre * IQ

                Df Sum of Sq RSS AIC
- class:pre:IQ  2      17.3      974 180
<none>                         957 183

Step:  AIC=180
post ~ class + pre + IQ + class:pre + class:IQ + pre:IQ

              Df Sum of Sq  RSS AIC
- class:IQ    2      22.9      997 177
- class:pre   2      37.7     1012 178
- pre:IQ      1      24.0      998 179
<none>                         974 180

Step:  AIC=177
post ~ class + pre + IQ + class:pre + pre:IQ

              Df Sum of Sq  RSS AIC
- class:pre   2      44.0     1041 176
<none>                         997 177
-pre:IQ      1      37.9     1035 177
```

```
Step:  AIC=176
post ~ class + pre + IQ + pre:IQ

           Df  Sum of Sq  RSS AIC
- pre:IQ   1       30      1071 175
<none>                     1041 176
- class    2      355      1396 188

Step:  AIC=175
post ~ class + pre + IQ

           Df Sum of Sq  RSS AIC
<none>                     1071 175
- IQ       1      176     1247 182
- class    2      334     1405 186
- pre      1      502     1574 195

Call:
lm(formula = post ~ class + pre + IQ, data = scores)

Coefficients:
(Intercept)      class2      class3         pre          IQ
    -14.876      -1.403       4.987       0.780       0.213
```

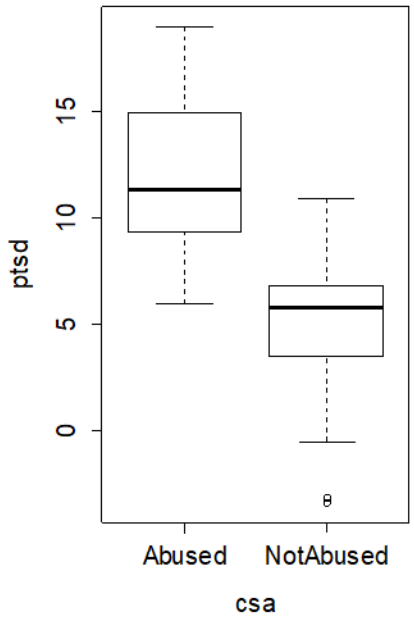
Variance partitioning table

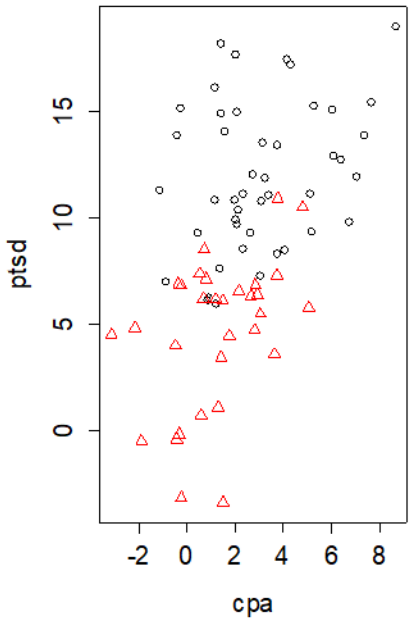
Source	DF	Type I SS	Mean Square	F Value	Pr > F
CLASSTYPE	2	115.6381579	57.8190789	2.75	0.0733
PRE	1	493.3922076	493.3922076	23.49	<.0001
IQ	1	175.7215915	175.7215915	8.36	0.0056

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CLASSTYPE	2	333.6317170	166.8158585	7.94	0.0010
PRE	1	502.1888091	502.1888091	23.91	<.0001
IQ	1	175.7215915	175.7215915	8.36	0.0056

Childhood sexual abuse

```
# Book: Linear models with R (Faraway  2009)
# Effects of childhood sexual abuse on adult females reported in
# Rodriguez et al. (1997):45 women treated at a clinic,
# who reported childhood sexual abuse (csa), were measured for
# post-traumatic stress disorder (ptsd) and
# childhood physical abuse (cpa)
```





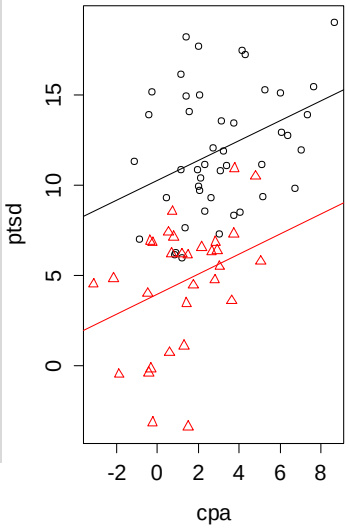

```
library(faraway)
data(sexab)
by(sexab, sexab$csa, summary)
plot(ptsd ~ csa, sexab)
plot(ptsd ~ cpa, pch = as.numeric(sexab$csa),
                 col = as.numeric(sexab$csa), sexab)
```

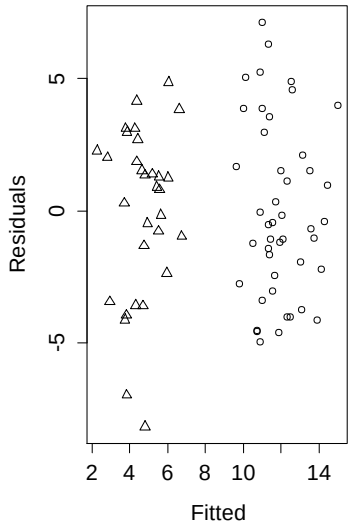
	cpa	ptsd	csa
1	2.04786	9.71365	Abused
2	0.83895	6.16933	Abused
3	-0.24139	15.15926	Abused
4	-1.11461	11.31277	Abused
5	2.01468	9.95384	Abused
6	6.71131	9.83884	Abused

```
m1 <- lm (ptsd ~ cpa+csa+cpa:csa, sexab)
summary (m1)
model.matrix (m1)

m2 <- lm (ptsd ~ cpa+csa, sexab)
summary (m2)
```

Childhood sexual abuse

```
Call:
lm(formula = ptsd ~ cpa + csa, data = sexab)
Residuals:
    Min      1Q  Median      3Q     Max
-8.1567 -2.3643 -0.1533  2.1466  7.1417
Coefficients:
            Estimate  Std. Error t value Pr(>|t|)
(Intercept)   3.9753    0.6293   6.317  1.87e-08 ***
cpa           0.5506    0.1716   3.209  0.00198  **
csaAbused     6.2728    0.8219   7.632  6.91e-11 ***
```



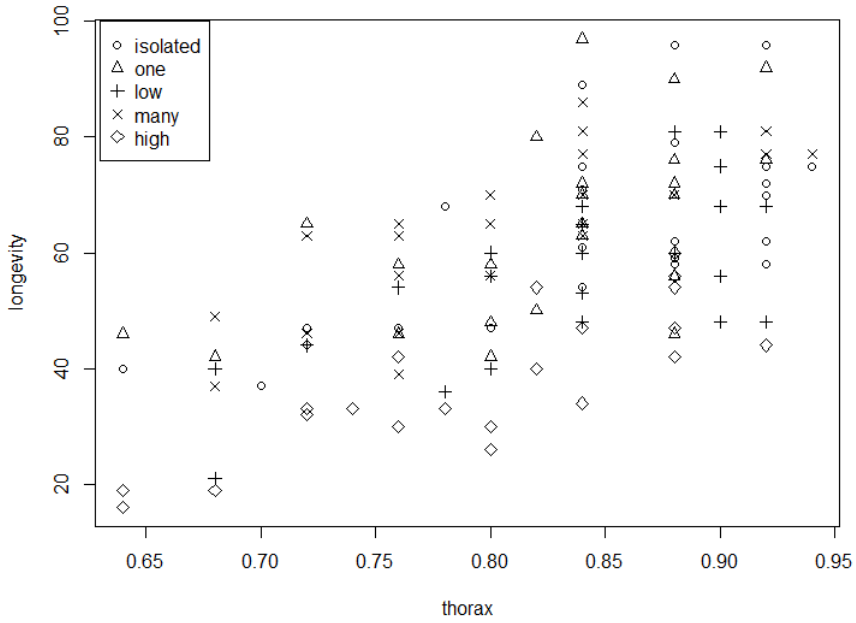



```
plot(ptsd~cpa, pch=as.numeric(sexab$csa),
      col=as.numeric(sexab$csa), sexab)
  abline (3.9753,  0.5506, col = 'red' ) # not abused
  abline (10.248,  0.5506) # abused, 10.248 = 3.9753 + 6.2728
plot (fitted (m2), residuals (m2), pch=as.numeric(sexab$csa),
       xlab= "Fitted", ylab="Residuals")

# change the reference level
sexab$csa <- relevel (sexab$csa, ref="NotAbused") # ref="Abused"
m3 <- lm (ptsd ~ cpa+csa, sexab)
summary (m3)
```

Sexual activity and the life span of male fruitflies

 # The data for this example come from a study on the sexual activity and the life span of
 # male fruitflies by Partridge and Farquhar (1981):125 fruitflies were divided randomly
 # into five groups of 25 each. The response was the longevity of the fruitfly in days. One
 # group was kept solitary, while another was kept individually with a virgin female each
 # day. Another group was given eight virgin females per day. As an additional control, the
 # fourth and fifth groups were kept with one or eight pregnant females per day. Pregnant
 # fruitflies will not mate. The thorax length of each male was measured as this was known
 # to affect longevity. The five groups are labeled many, isolated, one, low and high
 # respectively. The purpose of the analysis is to determine the difference between the five
 # groups if any.

```
library(faraway)
data (fruitfly)
plot (longevity ~ thorax, fruitfly, pch=unclass (activity))
legend (0.63, 100, levels (fruitfly$activity), pch=1:5)

g <- lm (longevity ~ thorax*activity, fruitfly)
summary (g)
model.matrix(g)
anova(g)

gb <- lm (longevity ~ thorax+activity, fruitfly)
drop1 (gb, test="F") # drop one term, using F test
```



Sexual activity and the life span of male fruitflies

summary (g)

```
lm(formula = longevity ~ thorax * activity, data = fruitfly)

Coefficients:
                    Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)         -50.2420    21.8012     -2.305    0.023 *
thorax              136.1268    25.9517      5.245    7.27e-07 ***
activityone           6.5172    33.8708      0.192    0.848
activitylow          -7.7501    33.9690     -0.228    0.820
activitymany         -1.1394    32.5298     -0.035    0.972
activityhigh        -11.0380    31.2866     -0.353    0.725
thorax:activityone   -4.6771    40.6518     -0.115    0.909
thorax:activitylow    0.8743    40.4253      0.022    0.983
thorax:activitymany   6.5478    39.3600      0.166    0.868
thorax:activityhigh -11.1268    38.1200     -0.292    0.771
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.71 on 114 degrees of freedom
Multiple R-squared:  0.6534,  Adjusted R-squared:  0.626
F-statistic: 23.88 on 9 and 114 DF,  p-value: < 2.2e-16
```

drop1 (gb, test="F")

```
Single term deletions

Model:
longevity ~ thorax + activity
          Df  Sum of Sq   RSS     AIC    F value   Pr(>F)
<none>                    13107   589.92
thorax     1   12368.4    25476   670.32  111.348  < 2.2e-16
activity   4    9634.6    22742   650.25   21.684  1.974e-13
```

Missing data

```
library(faraway)
data(chmiss) # Chicago insurance dataset
head(chmiss)
```

	race	fire	theft	age	involact	income
60626	10	6.2	29	60.4	NA	11.744
60640	22.2	9.5	44	76.5	0.1	9.323
60613	19.6	10.5	36	NA	1.2	9.948
60657	17.3	7.7	37	NA	0.5	10.656
60614	24.5	8.6	53	81.4	0.7	9.73
60610	54	34.1	68	52.6	0.3	8.231

```
model <- lm(involact ~ ., chmiss)
summary(model)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.116483 0.605761  -1.843  0.079475 .
race         0.010487 0.003128   3.352   0.003018 **
fire         0.043876 0.010319   4.252   0.000356 ***
theft       -0.017220 0.005900  -2.918   0.008215 **
age          0.009377 0.003494   2.684   0.013904 *
income       0.068701 0.042156   1.630   0.118077
Residual standard error: 0.3382 on 21 degrees of freedom (20 observations
deleted due to missingness) Multiple R-squared: 0.7911, Adjusted R-squared:
0.7414 F-statistic: 15.91 on 5 and 21 DF, p-value: 1.594e-06
```

Replacing missing data with mean

Any case with at least one missing value is omitted from the regression. There are now only 21 degrees of freedom - almost half the data is lost. We can fill in the missing values by their variable means as in:

```
cmeans <- apply (chmiss, 2, mean, na.rm=T)
cmeans
```

race	fire	theft	age	involact	income
35.60930	11.42444	32.65116	59.96905	0.64773	10.73587

```
mchm <- chmiss
for (i in c(1, 2, 3, 4, 6)) mchm[is.na (chmiss[,i]), i] <- cmeans[i]

model <- lm(involact ~ ., mchm)
summary(model)
```

Residual standard error: 0.3841 on 38 degrees of freedom **(3 observations deleted due to missingness)** Multiple R-squared: 0.682, Adjusted R-squared: 0.6401 F-statistic: 16.3 on 5 and 38 DF, p-value: 1.409e-08

Mixed effects ANCOVA

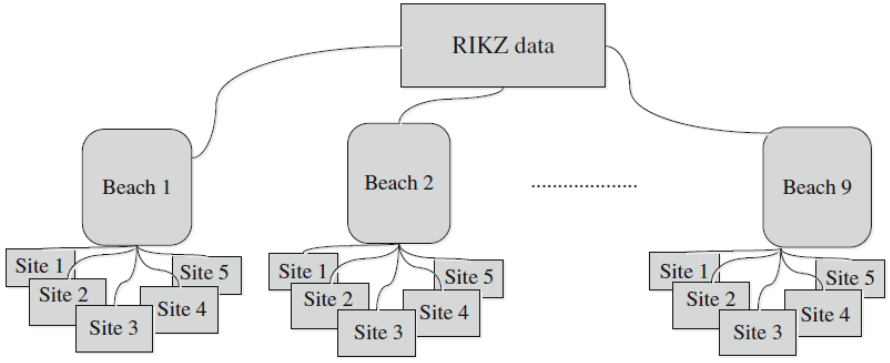
Sample	Richness	Exposure	NAP	Beach
1	11	10	0.045	1
2	10	10	-1.036	1
3	13	10	-1.336	1
4	11	10	0.616	1
5	10	10	-0.684	1
6	8	8	1.19	2
7	9	8	0.82	2
8	8	8	0.635	2
9	19	8	0.061	2
10	17	8	-1.334	2
11	6	11	-0.976	3
12	1	11	1.494	3
13	4	11	-0.201	3
14	3	11	-0.482	3
15	3	11	0.167	3
16	1	11	1.768	4
17	3	11	-0.03	4
18	3	11	0.46	4
19	1	11	1.367	4
20	4	11	-0.811	4
21	3	10	1.117	5
22	22	10	-0.503	5
23	6	10	0.729	5
24	0	10	1.627	5
25	6	10	0.054	5
26	5	11	-0.578	6
27	4	11	-0.348	6
28	1	11	2.222	6
29	6	11	-0.893	6
30	4	11	0.766	6
31	2	11	0.883	7
32	1	11	1.786	7
33	1	11	1.375	7
34	3	11	-0.06	7
35	4	11	0.367	7
36	3	10	1.671	8
37	5	10	-0.375	8
38	7	10	-1.005	8
39	5	10	0.17	8
40	0	10	2.052	8
41	7	10	-0.356	9
42	11	10	0.094	9
43	3	10	-0.002	9
44	0	10	2.255	9
45	2	10	0.865	9

Mixed effects modelling for ANCOVA

Marine benthic data were collected from nine inter-tidal areas along the Dutch coast by the Dutch institute RIKZ in the summer of 2002.

In each inter-tidal area (denoted by ‘beach’), five samples were taken, and the macro-fauna and abiotic variables were measured.

Species richness (the number of different species) can be explained by NAP (Normal Amsterdams Peil, the height of a sampling station compared to mean tidal level) and beaches.



Zuur et al. 2009. *Mixed Effects Models and Extensions in Ecology with R.* Page 101-102

The random intercept model

$$R_{ij} = \alpha + \beta_1 \times NAP_{ij} + \beta_2 \times Beach_i + \varepsilon_{ij}$$

Beach i

$$\begin{pmatrix} R_{i1} \\ R_{i2} \\ R_{i3} \\ R_{i4} \\ R_{i5} \end{pmatrix} = \begin{pmatrix} 1\ NAP_{i1} \\ 1\ NAP_{i2} \\ 1\ NAP_{i3} \\ 1\ NAP_{i4} \\ 1\ NAP_{i5} \end{pmatrix} \times \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \times b_i + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \\ \varepsilon_{i5} \end{pmatrix}$$

Assumptions:
The random effects b_i are normally distributed: N(0, d^2).
The errors $\boldsymbol{\varepsilon}_i$ are normally distributed.

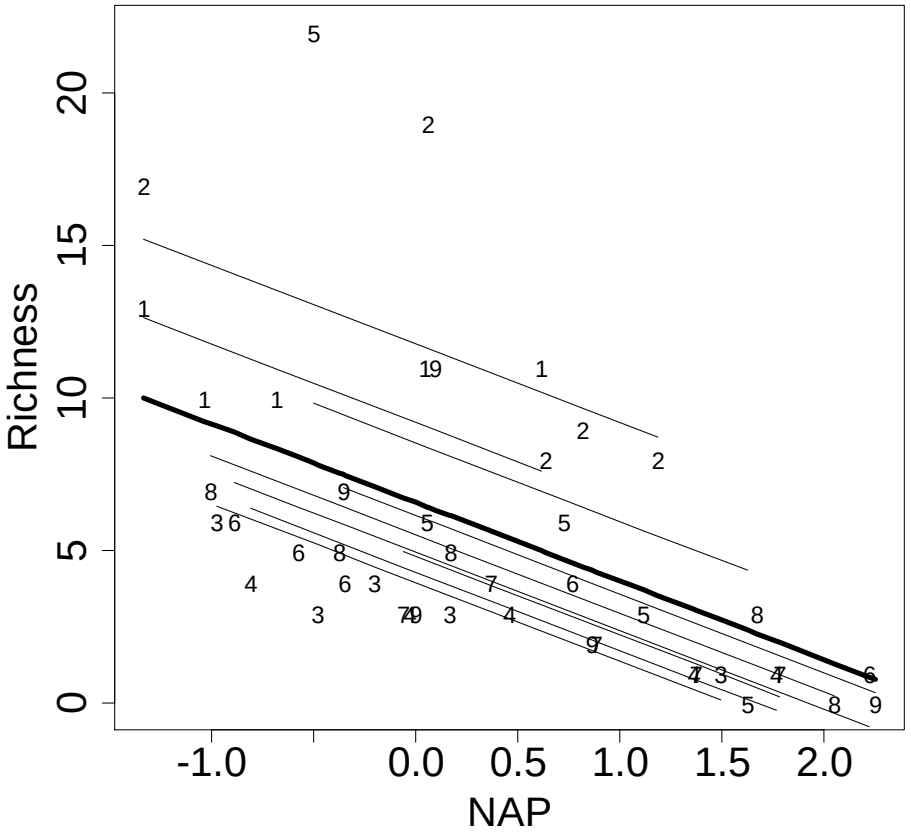
```
library(nlme)
RIKZ$fBeach <- factor(RIKZ$Beach)
Mlme1 <- lme(Richness ~ NAP, random = ~1 | fBeach, data = RIKZ)
summary(Mlme1)
```

Zuur et al. 2009. *Mixed Effects Models and Extensions in Ecology with R.* Page 106

The random intercept model: results

```
Linear mixed-effects model fit by REML
 Data: RIKZ
      AIC      BIC    logLik
 247.4802 254.525 -119.7401

Random effects:
 Formula: ~1 | fBeach
             (Intercept)  Residual
StdDev:    2.944065       3.05977

Fixed effects: Richness ~ NAP
                Value  Std.Error DF   t-value p-value
(Intercept)  6.581893  1.0957618 35  6.006682       0
NAP         -2.568400  0.4947246 35 -5.191574       0

Number of Observations: 45
Number of Groups: 9
```

The random intercept model: results



The thick line represents the fitted line obtained by the fixed component 6.58 – 2.56 NAP_i, also called the population model.

The other lines are obtained by adding the contribution of $\mathbf{b}_i$ for each beach i to the population fitted curve.

Hence, the random intercept model implies one average curve (the thick line) that is allowed to be shifted up, or down, for each beach by something that is normally distributed with a certain variance d^2 (2.94^2).

The random intercept and slope model

Beach i

$$\begin{pmatrix} R_{i1} \\ R_{i2} \\ R_{i3} \\ R_{i4} \\ R_{i5} \end{pmatrix} = \begin{pmatrix} 1 & NAP_{i1} \\ 1 & NAP_{i2} \\ 1 & NAP_{i3} \\ 1 & NAP_{i4} \\ 1 & NAP_{i5} \end{pmatrix} \times \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} 1 & NAP_{i1} \\ 1 & NAP_{i2} \\ 1 & NAP_{i3} \\ 1 & NAP_{i4} \\ 1 & NAP_{i5} \end{pmatrix} \times \begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \\ \varepsilon_{i5} \end{pmatrix}$$

Assumptions: $\begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \sim N(0, D)$ where $\mathrm{D} = \begin{pmatrix} d_{11}^{\ 2} & d_{21} \\ d_{12} & d_{22}^{\ 2} \end{pmatrix}$

The errors $\boldsymbol{\varepsilon}_i$ are normally distributed.

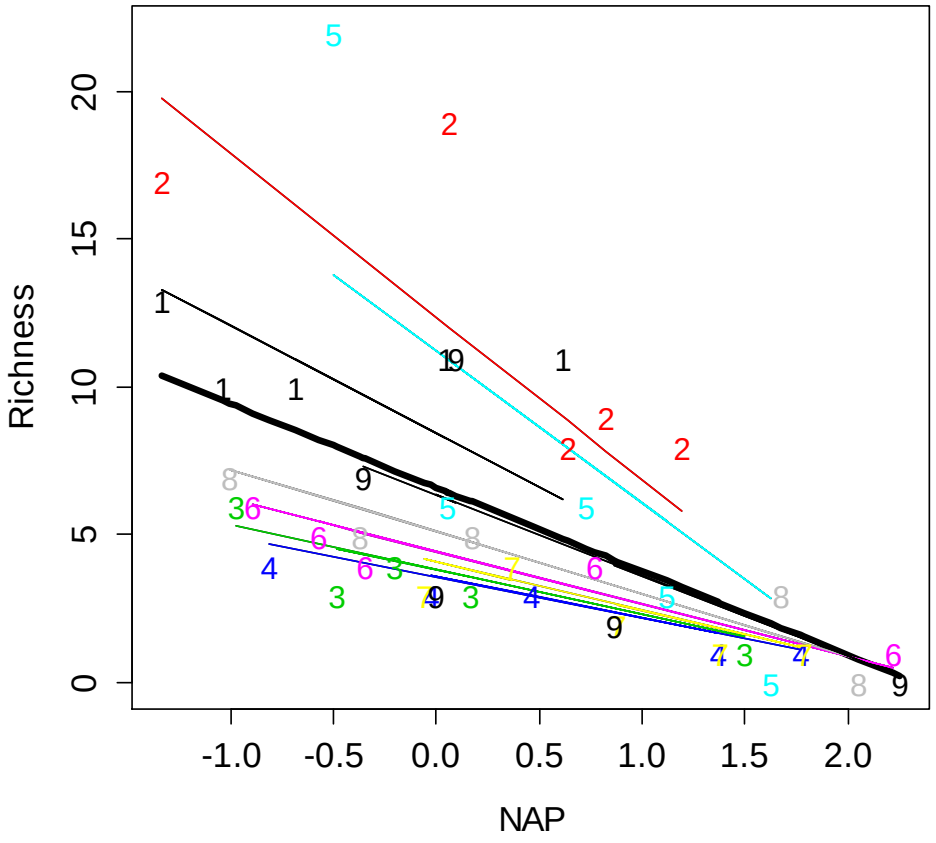
```
library(nlme)
RIKZ$fBeach <- factor(RIKZ$Beach)
Mlme2 <- lme(Richness ~ NAP, random = ~1 + NAP | fBeach, data = RIKZ)
summary(Mlme2)
```

Zuur et al. 2009. *Mixed Effects Models and Extensions in Ecology with R.* Page 109-110

The random intercept and slope model: results

```
Linear mixed-effects model fit by REML
 Data: RIKZ
      AIC      BIC    logLik
 244.3839 254.9511 -116.1919

Random effects:
 Formula: ~1 + NAP | fBeach
 Structure: General positive-definite, Log-Cholesky parametrization
             StdDev     Corr
(Intercept)  3.549100  (Intr)
NAP          1.715015  -0.99
Residual     2.702785

Fixed effects: Richness ~ NAP
                Value        Std.Error    DF    t-value      p-value
(Intercept)  6.588729    1.2647708   35    5.209425   0e+00
NAP         -2.830029    0.7229514   35   -3.914549   4e-04
Correlation:
(Intr) NAP -0.819

Number of Observations: 45
Number of Groups: 9
```

The random intercept and slope model



The variance d_{11}^2 plays the same role as d^2 in the random intercept model; it determines the amount of variation around the population intercept α. The numerical output shows that its estimated value is $3.54^2 = 12.5$.

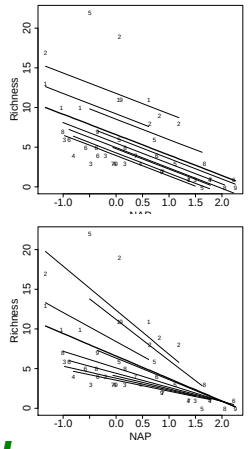
The model also allows for random variation around the population slope in a similar way as it does for the intercept. The variance d_{22}^2 determines the variation in slopes at the nine beaches. The estimated value of $1.71^2 = 2.92$ shows that there is considerably more variation in intercepts than in slopes at the nine beaches.

Finally, there is a correlation between the random intercepts and slopes. Its value of –0.99 is rather high, but indicates that beaches with a high positive intercept also have a high negative slope.

Zuur et al. 2009. *Mixed Effects Models and Extensions in Ecology with R.* Page 110

R code for figures

```
# The Random Intercept and/or slope Model
RIKZ = read.table('D:/softwares/R/library/AED/data/RIKZ.txt',header=T)
library(nlme)
RIKZ$fBeach <- factor(RIKZ$Beach)
Mlme1 <- lme(Richness ~ NAP, random = ~1 | fBeach, data = RIKZ)
Mlme1 <- lme(Richness ~ NAP, random = ~1 + NAP | fBeach, data = RIKZ)
summary(Mlme1)

# plot regression lines
F0 <- fitted(Mlme1, level = 0) # fitted values obtained by the population model
F1 <- fitted(Mlme1, level = 1) # fitted values obtained by within-beach model
I <- order(RIKZ$NAP); NAPs <- sort(RIKZ$NAP)
plot(NAPs, F0[I], lwd = 4, type = "l",
ylim = c(0, 22), ylab = "Richness", xlab = "NAP")
for (i in 1:9){
   x1 <- RIKZ$NAP[RIKZ$Beach == i]
   y1 <- F1[RIKZ$Beach == i]
   K <- order(x1)
   lines(sort(x1), y1[K])
}
text(RIKZ$NAP, RIKZ$Richness, RIKZ$Beach, cex = 0.9)
```

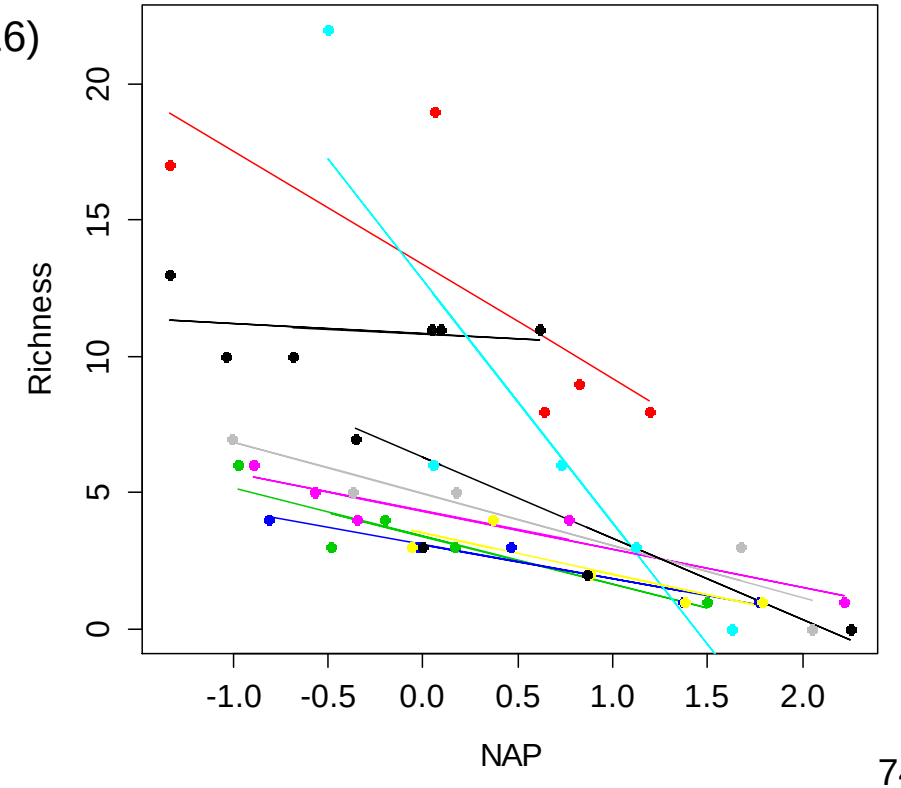


model = lm(Richness ~ NAP * fBeach, data = RIKZ)

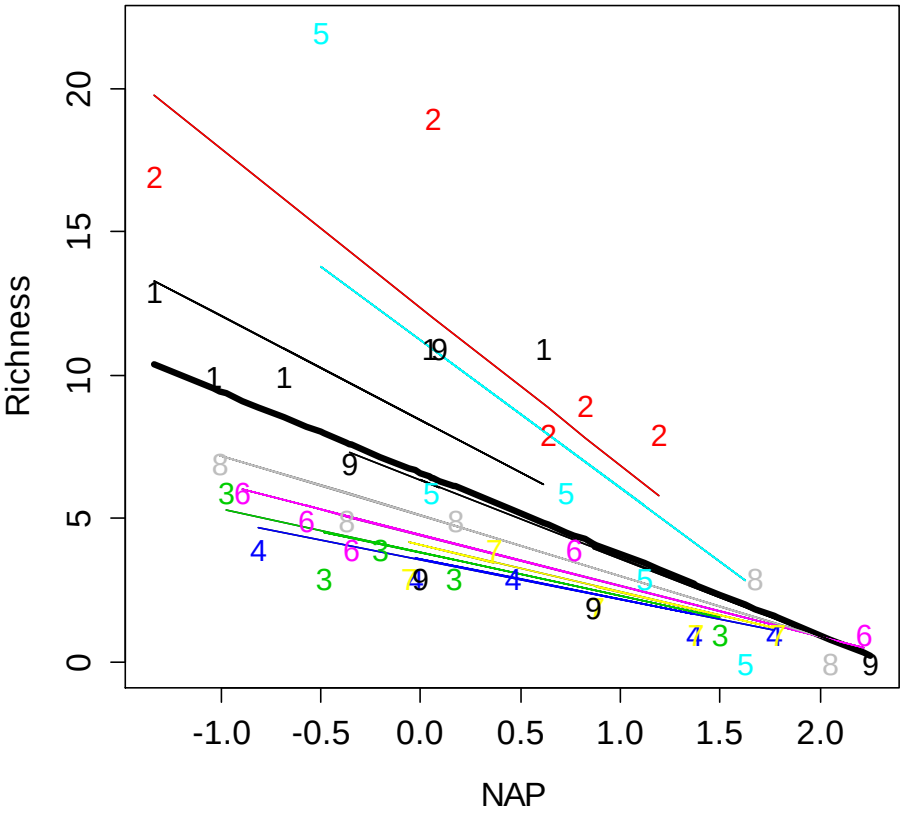
```
model = lm(Richness ~ NAP * fBeach, data = RIKZ)
anova(model)
pred = predict(model, RIKZ[, c('NAP', 'fBeach')])
plot(Dat$NAP, Dat$Richness, xlab='NAP', ylab='Richness', col='white')
Dat = cbind(RIKZ, pred)
for (i in 1:9) {
  Data = Dat[Dat$Beach==i, ]
  lines(Data$NAP, Data$pred, col=i)
  points(Data$NAP, Data$Richness, col=i, pch=16)
}
```

Coefficients:

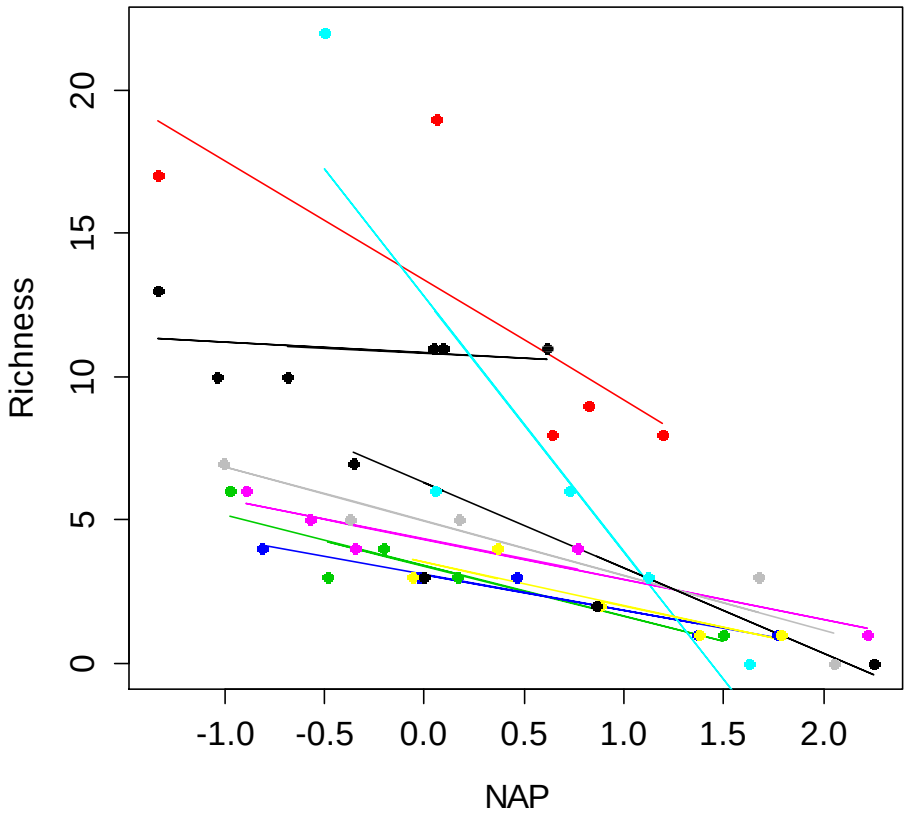
	Estimate	Std.Error	t value	p	
(Intercept)	10.8219	1.3341	8.112	1.03E-08	***
NAP	-0.3718	1.5493	-0.24	0.81214	
fBeach2	2.5238	1.7685	1.427	0.165013	
fBeach3	-7.4212	1.7346	-4.278	0.000211	***
fBeach4	-7.7342	1.8541	-4.171	0.000281	***
fBeach5	1.9609	1.9485	1.006	0.323166	
fBeach6	-6.4973	1.7494	-3.714	0.000938	***
fBeach7	-7.3013	2.2611	-3.229	0.003253	**
fBeach8	-5.8704	1.7981	-3.265	0.002974	**
fBeach9	-4.5268	1.8631	-2.43	0.022036	*
NAP:fBeach2	-3.8034	1.9941	-1.907	0.067169	.
NAP:fBeach3	-1.3835	2.0405	-0.678	0.503511	
NAP:fBeach4	-0.8767	1.9528	-0.449	0.657044	
NAP:fBeach5	-8.5283	2.134	-3.996	0.000447	***
NAP:fBeach6	-1.0167	1.829	-0.556	0.582872	
NAP:fBeach7	-1.1458	2.2756	-0.504	0.618683	
NAP:fBeach8	-1.5212	1.8134	-0.839	0.408891	
NAP:fBeach9	-2.5957	1.9537	-1.329	0.1951	



Mixed effect model vs. regular ANCOVA



lme(Richness ~ NAP, random = ~1 + NAP | fBeach, data = RIKZ)



lm(Richness ~ NAP * fBeach, data = RIKZ)

Rethinking the homogeneity of regression slopes assumption

Always run an ANCOVA model including both the IV, CV and the CVxIV interaction term.

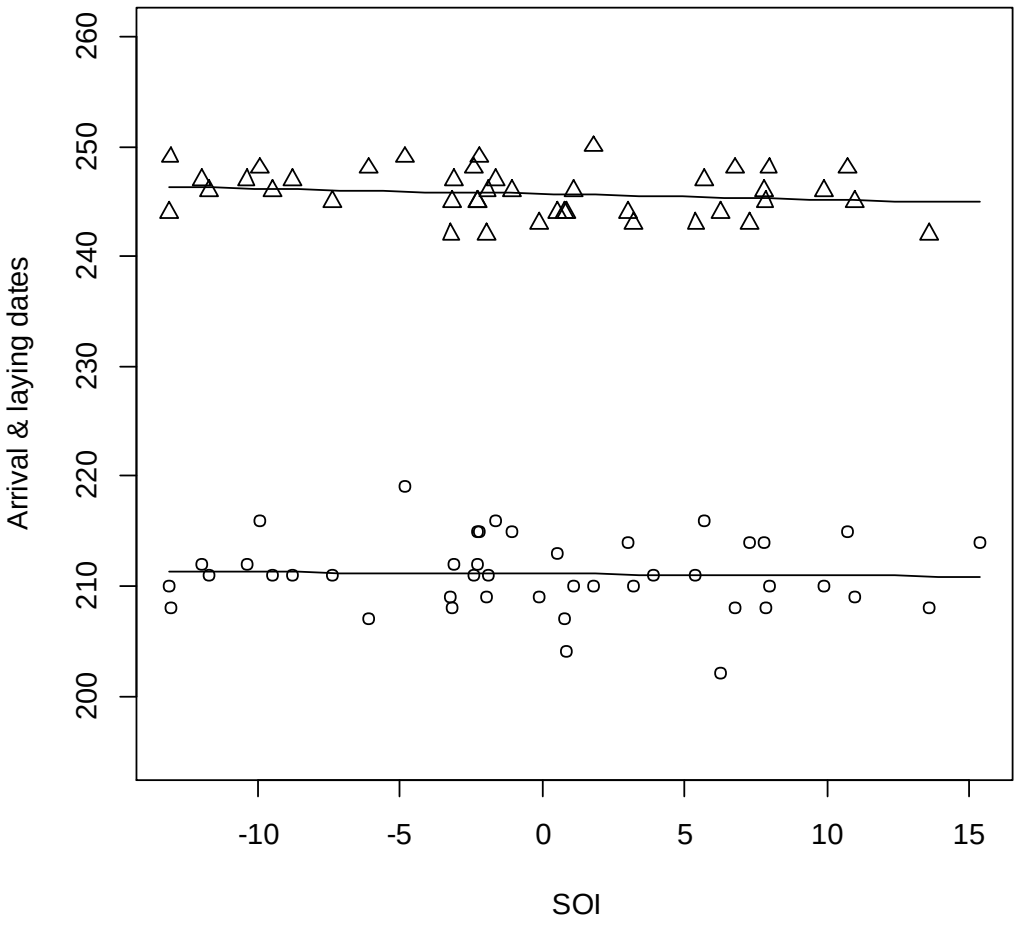
If the CVxIV interaction is significant, ANCOVA should not be performed.

One option is to assess group differences on the DV at particular levels of the CV.

Another option is to use mediation analyses to determine if the CV accounts for the IV's effect on the DV.

ANCOVA with autocorrelation

Year	ArrivalAP	LayingAP	ArrivalCP	LayingCP	ArrivalEP	LayingEP	MSA	SOI
1951	214	NA	NA	NA	NA	NA	0.086	15.38
1952	NA	NA	NA	NA	-10	46	0.127	-0.69
1953	212	245	209	260	NA	NA	0.101	-2.28
1954	NA	NA	NA	NA	NA	NA	0.109	-6.8
1955	NA	NA	NA	NA	NA	NA	0.087	4.08
1956	NA	NA	NA	NA	-9	42	0.086	10.58
1957	215	248	214	259	NA	NA	0.072	10.73
1958	NA	NA	NA	NA	-13	44	0.082	-3.89
1959	209	242	NA	NA	-3	38	0.078	-3.2
1960	NA	NA	NA	NA	NA	NA	0.078	-0.04
1961	NA	NA	NA	NA	NA	NA	0.065	3.83
1962	207	244	210	NA	-12	NA	0.086	0.8
1963	211	243	193	261	-7	41	0.137	5.4
1964	209	242	213	NA	-15	43	0.098	-1.95
1965	202	244	207	262	NA	NA	0.111	6.28
1966	NA	NA	NA	NA	-7	41	0.098	-8.43
1967	NA	NA	NA	NA	-9	38	0.056	-4.24
1968	210	243	212	260	-17	44	0.065	3.2
1969	214	244	209	NA	NA	NA	0.082	3.02
1970	NA	NA	NA	NA	-1	46	0.067	-5.38
1971	211	NA	207	NA	-6	45	0.051	3.93
1972	209	245	201	261	0	44	0.064	10.95
1973	211	245	215	268	0	50	0.071	-7.35
1974	214	243	NA	NA	-7	44	0.063	7.28
1975	210	246	195	259	-7	44	0.106	9.9
1976	208	242	198	262	-15	44	0.087	13.6
1977	210	246	210	266	-1	43	0.078	1.11
1978	216	248	213	265	-5	43	0.064	-9.9
1979	216	247	213	265	-9	43	0.08	-1.65
1980	211	246	206	NA	-11	51	0.037	-1.91
1981	212	247	213	NA	-5	45	0.054	-3.08
1982	210	250	212	NA	-20	46	0.06	1.8
1983	208	249	216	264	NA	NA	0.09	-13.05
1984	NA	NA	NA	NA	-10	39	0.088	-8.33
1985	209	243	208	260	-4	42	0.082	-0.11
1986	204	244	199	263	-10	45	0.061	0.86
1987	211	248	210	261	-4	44	0.096	-2.38
1988	210	244	213	263	-4	46	0.048	-13.08
1989	208	245	220	NA	-1	47	0.051	7.82
1990	208	248	219	265	-2	47	0.034	6.77
1991	215	249	215	NA	-3	47	0.054	-2.19
1992	211	247	211	264	-6	44	0.041	-8.78
1993	212	247	209	269	-7	41	0.056	-10.38
1994	211	246	215	262	-7	41	0.076	-9.47
1995	212	247	215	263	-17	47	0.065	-11.93
1996	215	245	215	263	-13	46	0.054	-2.27
1997	216	247	214	NA	-6	42	NA	5.69
1998	211	246	217	262	-1	NA	NA	-11.67
1999	215	246	220	263	-4	47	NA	-1.08
2000	210	248	208	261	-3	43	NA	7.95
2001	214	246	NA	263	-14	47	NA	7.8
2002	213	244	215	NA	-16	44	NA	0.53
2003	207	248	NA	264	-9	43	NA	-6.1
2004	208	245	216	265	-11	45	NA	-3.14
2005	219	249	210	266	NA	NA	NA	-4.82



Zuur et al. 2009. *Mixed Effects Models and Extensions in Ecology with R.* Page 343-361

ANCOVA with autocorrelated data

```
head(ABirds) # data

 AP    <- c(ABirds$ArrivalAP, ABirds$LayingAP)
 SOI2 <- c(ABirds$SOI, ABirds$SOI)
 Y2    <- c(ABirds$Year, ABirds$Year)
 ID    <- factor(rep(c("Arrival", "Laying"), each = 55))

 library(nlme)
 vf2 <- varIdent(form =~ 1 | ID)
 M1 <- gls(AP ~ SOI2 + ID + SOI2:ID, weights = vf2, na.action = na.omit)
 M2 <- gls(AP ~ SOI2 + ID + SOI2:ID, weights = vf2, na.action = na.omit,
           correlation = corAR1(form =~Y2 | ID))
 anova(M1, M2)
```

```
   Model df      AIC      BIC    logLik   Test   L.Ratio p-value
M1     1  6 427.8205 442.2608 -207.9102                        
M2     2  7 426.0757 442.9228 -206.0379 1 vs 2  3.744727   0.053
```

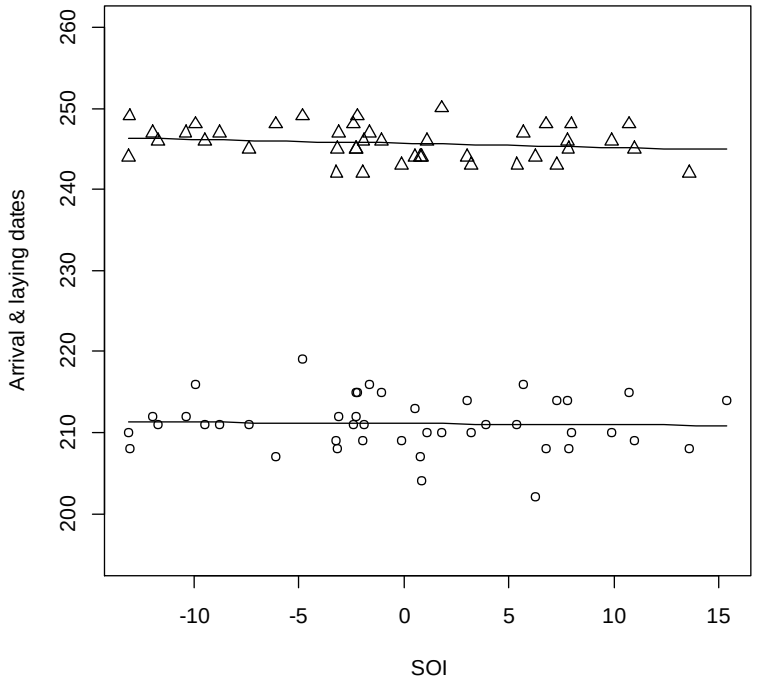
Model selection

```
# to compare two models with the same random structure, but with different fixed effect,
# we need to use the maximum likelihood estimation method instead of REML
# (P.356 in Zuur et al. 2009)
M3 <- gls(AP ~ SOI2 + ID + SOI2:ID, weights = vf2, na.action = na.omit, method = "ML",
          correlation = corAR1(form =~Y2 | ID))
M4 <- gls(AP ~ SOI2 + ID, weights = vf2, na.action = na.omit, method = "ML",
          correlation = corAR1(form =~Y2 | ID))
M5 <- gls(AP ~ ID, weights = vf2, na.action = na.omit, method = "ML",
          correlation = corAR1(form =~Y2 | ID))

anova(M3, M4, M5)
```

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
M3	1	7	419.0303	436.2107	-202.5152			
M4	2	6	417.1962	431.9222	-202.5981	1 vs 2	0.1658611	0.6838
M5	3	5	416.2205	428.4922	-203.1102	2 vs 3	1.0243171	0.3115

Plot

```
plot(ABirds$SOI, ABirds$ArrivalAP, ylim = c(195, 260), type = "n",
     ylab = "Arrival & laying dates", xlab='SOI')
points(ABirds$SOI, ABirds$ArrivalAP, pch = 1)
points(ABirds$SOI, ABirds$LayingAP, pch = 2)
MyX <- data.frame(SOI2 = seq(from = min(ABirds$SOI),
          to = max(ABirds$SOI),
          length = 20), ID = "Arrival")
Pred1 <- predict(M3, newdata = MyX)
lines(MyX$SOI2, Pred1)
MyX <- data.frame(SOI2 = seq(from = min(ABirds$SOI),
          to = max(ABirds$SOI),
          length = 20), ID = "Laying")
Pred2 <- predict(M3, newdata = MyX)
lines(MyX$SOI2, Pred2)
```



Model formulas in R

Regression	y~x	x is a continuous explanatory variable
One-way ANOVA	y~type	type is a factor (categorical variable)
Random block design	y~type+treatment	type and treatment are factors
Factorial ANOVA	y~n*p*k	Include main effect and all interaction
Three-way ANOVA	y~n*p*k-n:p:k	As above, don't fit three-way interaction
Analysis of covariance	y~x+type	Common slope but different intercepts
Analysis of covariance	y~x*type	Different slopes and different intercepts
Nested ANOVA	y~a/b/c	Factor c nested in factor b, within factor a
Split-plot ANOVA	y~a/b/c+Error(a/b/c)	Factorial experiment with three different error terms
Multiple regression	y~x+z	Two continuous explanatory variables
Multiple regression	y~x*z	Includes interaction: x+y+x:y
Multiple regression	y~x+I(x^2)+z +I(z^2)+x:z	Quadratic. I() indicates as is, so I(x^2) is x squared.
Multiple regression	y~poly(x,2)	Quadratic polynomial
Multiple regression	y~(w+x+z)^2	Fit variable and their interactions up to two-way
Nonparametric model	y~s(x)+s(z)	Fit smoothed x an z in a generalized additive model
Transformed response & explanatory variables	log(y)~I(1/x)+sqrt(z)	Transformation specified in model

Model Operators

Symbol	Explanation
+	indicates inclusion of an explanatory variable, not addition
-	indicates deletion of an explanatory variable, not subtraction
*****	indicates inclusion of explanatory variables and all their interactions, not multiplication
/	indicates nesting of explanatory variable, not division
|	indicates conditioning
:	indicates an interaction, such as x:z
a*b*c	= a + b + c + a:b + a:c + b:c +a:b:c
a/b/c	= a + b%in%a + c%in%b%in%a
(a+b+c)^2	= a + b + c + a:b + a:c + b:c = main effects & up to 2-way interactions
a*b*c-a:b:c	= a + b + c + a:b + a:c + b:c

Statistical Methods

Method	Description
lm	Fits a linear model with normal errors and constant variance: regression, analysis of variance, analysis of covariance
aov	Also fits a linear model with normal errors and constant variance, but oriented towards analysis of variance
glm	Fits generalized linear models by specifying one of a family of error structures (e.g., Poisson for count data) and a particular link function.
gam	Fits generalized additive models by specifying one of a family of error structures (e.g., Poisson for count data) in which continuous explanatory variable can optionally be fitted as arbitrary smoothed functions using non-parametric smoothers rather than a specific parametric functions.
lme & lmer	Fits linear mixed-effects models including fixed and random effects and allow for the specification of correlation structure amongst the explanatory variables and autocorrelation of the response variable. lmer allows for non-normal errors and non-constant variance with the same error families as glm
nls	Fits non-linear regression models using least squares.
loess	Fits a local regression model with one or more continuous explanatory variables using non-parametric techniques to produce a smoothed model surface.
tree	fits a regression tree model using binary recursive partitioning.

Functions

Functions	Description
summary	Displays the results of a model fit, depending on the method used. Use summary.lm or summary.aov to obtain particular displays.
anova	Displays an analysis of variance table and compares models.
plot	Produces diagnostic plots for model checking.
update	Modifies the last model fit.
coef	Displays the estimated coefficients from a model.
fitted	Displays the fitted values from a model.
resid	Displays the residuals from a model.
predict	Displays the predicted values from a model.

```
ancova = lm(Y~X1*X2)

coef(ancova)
fitted(ancova)
predict(ancova)
resid(ancova)
```

Assignment

Task:

Develop a ANCOVA experimental design. Generate your own data and FORMALIZE your hypotheses.

Define dependent variable, treatment variable and covariate variable

Check all assumptions for ANCOVA (including homogeneity of regression).

Clearly state the model performance, list the model fit statistics.