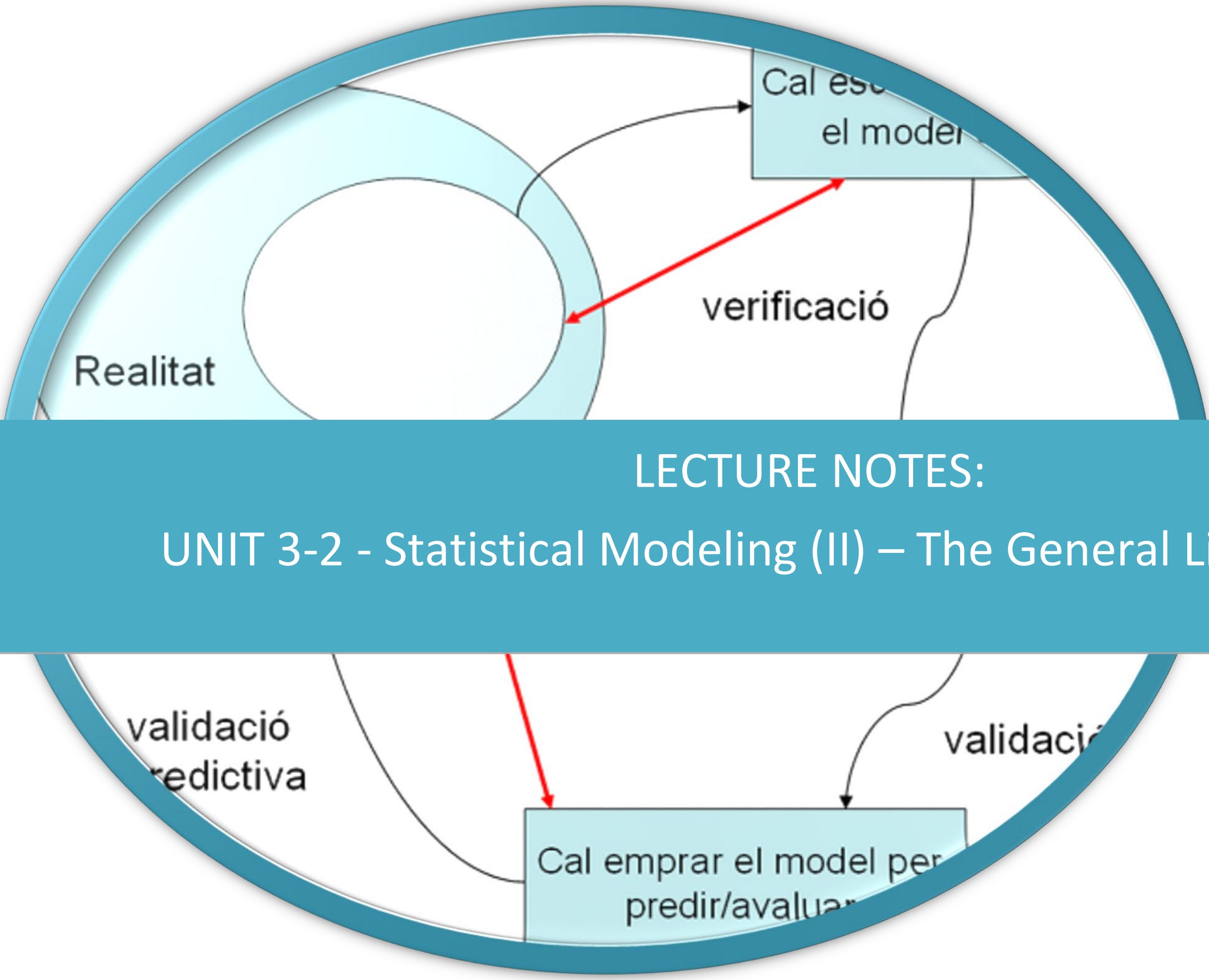
2022-2023 Academic year

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa



SIM course. Master in Data Science – FIB-UPC

LECTURE NOTES:

UNIT 3-2 - Statistical Modeling (II) – The General Linear Model

CONTENTS

3.2-1 READING LIST

Basic references:

📖 Fox, J. *Applied Regression Analysis and Generalized Linear Models*. Sage Publications, 2nd Edition 2008.
📖 Fox and Weisberg. ***An R Companion to Applied Regression***. Sage Publications, 2nd Edition 2010.

3.2-2 INTRODUCTION TO THE GENERAL LINEAR MODEL

Let $\mathbf{y}^T = (y_1, \ldots, y_n)$ be a vector of *n* observations, randomly drawn from the vector $\mathbf{Y}^T = (Y_1, \ldots, Y_n)$, whose variables are statistically independent and distributed with expectation $\boldsymbol{\mu}^T = (\mu_1, \ldots, \mu_n)$:

In linear models, the **random component** $\mathbf{Y}^T = (Y_1, \ldots, Y_n)$ is assumed to be normally distributed $Y_i|X_i \sim N(\mu_i, \sigma)$ with constant variance σ^2, $V(Y_i|X_i) = \sigma^2$ and expectation $E(Y_i|X_i) = \mu_i$

- Therefore, the response variable is modeled as normally distributed; thus, negative or positive values, which may be arbitrarily small or large, may be encountered as data for the response and prediction.
- **The systematic component** of the model consists in specifying a vector called the linear predictor, denoted η, of the same length as the response, dimension *n*. In vector notation, the parameters are $\boxed{\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)}$ and the regressors are $\mathbf{X} = (X_1, \ldots, X_p)$ and, thus, $\boxed{\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}}$ where η is *nx1*, $\mathbf{X}$ is *nxp* and β is *px1*.
- **Now, we will learn how to include factors as explanatory variables in the model.**
- Vector $\boldsymbol{\mu}$ is the direct linear predictor η; therefore, the **link function** is $\boldsymbol{\eta} = \boldsymbol{\mu}$ or $\mu_i = X_i\beta$

3.2-2. INTRODUCTION TO THE GENERAL LINEAR MODEL

The general linear model is an extension of multiple regression models to deal with explanatory variables as factors (nominal or qualitative variables) and the interaction between covariates (numeric variables) and factors.

- Linear regression can be extended to accommodate categorical variables (factors) using dummy variables (or indicator variables).

- ANOVA models can have one, two or more factors in the linear predictor, leading to one-way ANOVA, two-way ANOVA and general ANOVA. When factors produce additive effects of the first order, second order etc., interaction effects may appear.

- ANCOVA models combine factors (either main effects and/or interactions), covariates and interactions between factors (groups of factors) and covariates.

We are going to see:

1. One-way ANOVA: One factor and main effects only.
2. Two-way ANOVA: Main effects and interaction effects occur.
3. K Way ANOVA: Extension to multiple factors and high-order interactions.
4. ANCOVA models: One factor and one covariate.
5. Extension to multiple factors, covariates and interactions: the general linear model.

3.2-2. INTRODUCTION TO THE GENERAL LINEAR MODEL

- General linear models are estimated by least squares estimation.
- Model validation can be assessed using standard techniques of residual analysis. Unusual and influential data can be detected using standard techniques.
- Be careful when interpreting t-tests in ordinary output tables for regression.
- Interpretation and inference can be misleading when there are high-order interactions.
- It doesn't make much sense to standardize dummy variables:
  - They cannot be increased by standard deviation, so the regular interpretation for standardized coefficients does not apply.
  - Moreover, the standard interpretation of dummy variables showing differences in level between two categories is lost.
  - We cannot standardize interaction effects; they are dependent on the main effects.
  - We can, however, standardize quantitative variables beforehand and construct interaction terms afterwards.

3.2-3 ONE-WAY ANOVA

- Does height depend on gender?
- Does profession prestige in the Duncan data depend on the type of profession?

Both height and prestige are numeric response data and we can assume there is a first random component stated as normal leading to an OLS estimator.

- Gender is a dichotomous factor (two levels: male or female).
- Type of profession is a polytomous factor, consisting of three levels: "blue collar" "white collar" and "professional".

How do we interpret the R^2 ? Just as we did in multiple regression:

- A high R^2 means that the regressors explain the variation in Y.
- A high R^2 does not mean that you have eliminated omitted-variable bias.
- A high R^2 does not mean that the variables included are statistically significant; this must be determined using hypothesis tests.

3.2-3. ONE-WAY ANOVA

The ANOVA model for one factor (usually with I level) - Grasping the basic concepts:

- Formulation and construction of the design matrix for the models of regression.
- Interpretation of its parameters.
- Discussion of inference.

Group 1	$y_{11}, y_{12}, \cdots, y_{1n_1}$	Mean $\bar{y}_1$
Group 2	$y_{21}, y_{22}, \cdots, y_{2n_2}$	Mean $\bar{y}_2$
...	...	...
Group I	$y_{I1}, y_{I2}, \cdots, y_{In_I}$	Mean $\bar{y}_I$

(1) $Y_{ij} = \mu_i + \varepsilon_{ij}$, I parameters $\boldsymbol{\varepsilon} \approx \mathbf{N}_n\left(\mathbf{0}, \sigma^2 \mathbf{I}\right)$.

(2) $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, μ is the overall expected mean and α_i effect for *i* level, I+1 parameters.

The usual null hypothesis is that there are no differences between the expected mean of the groups, which can be written according to the formulas as:

(1) $\mathbf{H_0}: \mu_1 = \cdots = \mu_I = \mu$ versus $\mathbf{H_1}: \exists \mu_i \neq \mu$.

(2) $\mathbf{H_0}: \alpha_1 = \cdots = \alpha_I = 0$ versus $\mathbf{H_1}: \exists \alpha_i \neq 0$.

3.2-3. ONE-WAY ANOVA

To simplify, assume without loss of generality that group sizes are equal to J, $n_i = J \quad i = 1, \ldots, I$

Formula (1) $Y_{ij} = \mu_i + \varepsilon_{ij}$

$$\mathbf{Y} = \begin{array}{c} 1 \\ \vdots \\ I \end{array} \begin{pmatrix} \left\{ \begin{matrix} y_{11} \\ \vdots \\ y_{1J} \end{matrix} \right. \\ - \\ \vdots \\ - \\ \left\{ \begin{matrix} y_{I1} \\ \vdots \\ y_{IJ} \end{matrix} \right. \end{pmatrix}, \quad \mathbf{X} = \begin{array}{c} \begin{matrix} 1 & 2 & & I \end{matrix} \\ \begin{pmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{pmatrix} \end{array}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_I \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \left\{ \begin{matrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1J} \end{matrix} \right. \\ - \\ \vdots \\ - \\ \left\{ \begin{matrix} \varepsilon_{I1} \\ \vdots \\ \varepsilon_{IJ} \end{matrix} \right. \end{pmatrix}$$

$$\boxed{\mathbf{b} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \begin{pmatrix} J & & 0 \\ & \ddots & \\ 0 & & J \end{pmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^{J} y_{1j} \\ \vdots \\ \sum_{j=1}^{J} y_{Ij} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_I \end{pmatrix}}$$

Therefore, the estimator of the parameters is the average of the groups, assuming the number of replies per class is identical and equal to (J).

The disadvantage of this formulation is that it cannot be easily extended to more than one factor and therefore the ANOVA formula (2) is generally used.

3.2-3. ONE-WAY ANOVA

Formula (2) $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ has I+1 parameters and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is a singular matrix

$$
\mathbf{Y} = \begin{array}{c} 1 \\ \vdots \\ I \end{array} \begin{pmatrix} \left\{ \begin{matrix} y_{11} \\ \vdots \\ y_{1J} \end{matrix} \right. \\ - \\ \vdots \\ - \\ \left\{ \begin{matrix} y_{I1} \\ \vdots \\ y_{IJ} \end{matrix} \right. \end{pmatrix}, \quad
\mathbf{X} = \begin{array}{c} \begin{matrix} \overbrace{}^{1} & \overbrace{}^{i=1} & \overbrace{}^{i=2} & \overbrace{}^{\cdots} & \overbrace{}^{i=I} \end{matrix} \\ \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \vdots & \mathbf{0} \\ \mathbf{1} & \vdots & \mathbf{0} & \vdots & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \vdots & \cdots & \vdots \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{pmatrix} \end{array}, \quad
\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_I \end{pmatrix}, \quad
\boldsymbol{\varepsilon} = \begin{pmatrix} \left\{ \begin{matrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1J} \end{matrix} \right. \\ - \\ \vdots \\ - \\ \left\{ \begin{matrix} \varepsilon_{I1} \\ \vdots \\ \varepsilon_{IJ} \end{matrix} \right. \end{pmatrix}, \quad
\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{pmatrix} n & J & \cdots & J \\ J & J & 0 & \\ \vdots & 0 & \ddots & 0 \\ J & 0 & 0 & J \end{pmatrix}
$$

There is no single solution to the normal equations, but rather infinite solutions and all of them give a squared sum of residuals of equal value.

Technically, there are endless possibilities for formulating an equivalent regression model, but with a single solution it is enough to add any restriction of the type $\omega_0 \mu + \sum_{i=1}^{I} \omega_i \alpha_i = 0$.

The most frequently used reparameterizations are baseline and zero-sum.

3.2-3. ONE-WAY ANOVA

Formula (2) $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ and the restriction $\alpha_1 = 0$

$$\mathbf{Y} = \begin{array}{c} 1 \\ \vdots \\ I \end{array} \begin{pmatrix} \left\{ \begin{array}{c} y_{11} \\ \vdots \\ y_{1J} \end{array} \right. \\ - \\ \vdots \\ - \\ \left\{ \begin{array}{c} y_{I1} \\ \vdots \\ y_{IJ} \end{array} \right. \end{pmatrix}, \quad \mathbf{X_R} = \underset{n \times I}{\begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & 1 & 0 & \vdots \\ \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \cdots & 1 \end{pmatrix}}, \quad \boldsymbol{\beta}_{P} = \begin{pmatrix} \mu \\ \alpha_2 \\ \vdots \\ \alpha_I \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \left\{ \begin{array}{c} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1J} \end{array} \right. \\ - \\ \vdots \\ - \\ \left\{ \begin{array}{c} \varepsilon_{I1} \\ \vdots \\ \varepsilon_{IJ} \end{array} \right. \end{pmatrix}$$

$$\mathbf{b}_R = \left(\mathbf{X_R}^{\mathrm{T}} \mathbf{X_R}\right)^{-1} \mathbf{X_R}^{\mathrm{T}} \mathbf{Y} = \begin{pmatrix} n & J & \cdots & J \\ J & J & & 0 \\ \vdots & \vdots & \ddots & \\ J & 0 & \cdots & J \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij} \\ \sum_{j=1}^{J} y_{1j} \\ \vdots \\ \sum_{j=1}^{J} y_{I-1,j} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 - \bar{y}_1 \\ \vdots \\ \bar{y}_I - \bar{y}_1 \end{pmatrix}$$

- ➔ I parameters.
- ➔ The average estimate for baseline group 1 is contained in the estimate of the first parameter μ and the additive effect for *i group-level in relation to the baseline group is the estimate of* α_i .
- ➔ Design matrices for treatment contrast reparameterization according to R project terminology can be shown with the method contr.treatment(I).

3.2-3. ONE-WAY ANOVA

Formula (2) $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ and the restriction $\sum_{i=1}^{I} \alpha_i = 0$ (o $\alpha_I = -\sum_{i=1}^{I-1} \alpha_i$): the overall mean is the first parameter estimate and the additive group-level effect over the mean response is estimated through α_i.

$$
\mathbf{Y} = \begin{array}{c} 1 \\ \vdots \\ I \end{array} \begin{pmatrix} \begin{Bmatrix} y_{11} \\ \vdots \\ y_{1J} \end{Bmatrix} \\ - \\ \vdots \\ - \\ \begin{Bmatrix} y_{I1} \\ \vdots \\ y_{IJ} \end{Bmatrix} \end{pmatrix}, \quad \mathbf{X_R} = \underset{n \times I}{\begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \vdots \\ \vdots & \vdots & \ddots & \mathbf{1} \\ \mathbf{1} & -\mathbf{1} & \cdots & -\mathbf{1} \end{pmatrix}} \quad \boldsymbol{\beta}_R = \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_{I-1} \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \begin{Bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1J} \end{Bmatrix} \\ - \\ \vdots \\ - \\ \begin{Bmatrix} \varepsilon_{I1} \\ \vdots \\ \varepsilon_{IJ} \end{Bmatrix} \end{pmatrix}, \quad \mathbf{b}_R = \left(\mathbf{X}_R^{\mathbf{T}} \mathbf{X}_R\right)^{-1} \mathbf{X}_R^{\mathbf{T}} \mathbf{Y} == \begin{pmatrix} \bar{y} \\ \bar{y}_1 - \bar{y} \\ \vdots \\ \bar{y}_{I-1} - \bar{y} \end{pmatrix}
$$

- Only I-1 additive effects are considered since the last level effect on the overall mean is minus the sum of the effects of levels 1 to I-1.

3.2-3. ONE-WAY ANOVA

➔ The number of independent parameters is *I*. Reduced design matrices $\mathbf{X}_R^{\mathrm{T}}\mathbf{X}_R$ are nonsingular *IxI*. The columns in design matrices are *dummies or dummy variables* denoted $D_1, \cdots, D_{I-1}$.

➔ Estimates must be interpreted as

$$\hat{\mu} = \frac{\sum_{i=1}^{I} \hat{\mu}_i}{I}, \quad \alpha_i = \hat{\mu}_i - \hat{\mu} \text{ and } \alpha_I = -\sum_{i=1}^{I-1} \alpha_i \text{ where } \hat{y}_{ij} = \overline{y} + \alpha_i = \overline{y}_i .$$

➔ Design matrices for sum contrast reparameterization according to R project terminology can be shown with the method contr.sum(I).

```
> contr.sum(3)
  [,1] [,2]
1    1    0
2    0    1
3   -1   -1
> contr.sum(4)
  [,1] [,2] [,3]
1    1    0    0
2    0    1    0
3    0    0    1
4   -1   -1   -1
```

```
> contr.treatment(3)
  2 3
1 0 0
2 1 0
3 0 1
> contr.treatment(4)
  2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
4 0 0 1
```

3.2-3. ONE-WAY ANOVA

➔ Comparing the null hypothesis $\mathbf{H_0}: \alpha_1 = \cdots = \alpha_I = 0$ to $\mathbf{H_1}: \exists \alpha_i \neq 0$ in (Formula 2) $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ plus either zero-sum or baseline reparameterization; the conclusion should be the same model interpretation given to different group means and must be taken into account by the statistician.

If **H₁** is true, the residual sum of squared *RSS₁*, for model $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ satisfies $RSS_1 / \sigma^2 \approx \chi^2_{n-I}$.

And provided $\mathbf{H_0}: \alpha_1 = \cdots = \alpha_I = 0$ is true, then $RSS_o = TSS = \sum\sum \left(y_{ij} - \bar{y}\right)^2$, $RSS_0 / \sigma^2 \approx \chi^2_{n-1}$ $RSS_0 - RSS_1 / \sigma^2 \approx \chi^2_{I-1}$ and F-test

$$f = \frac{RSS_0 - RSS_1}{I-1} \bigg/ \frac{RSS_1}{n-I} \approx \mathbf{F}_{I-1,\, n-I}$$

3.2-3. ONE-WAY ANOVA: EXAMPLE

Example: Prestige of Canadian occupations in data.frame Prestige in car library for R (Fox and Weisberg 2011)

- Description: The Prestige data frame has 102 rows and 6 columns. The observations are occupations. This data frame contains the following columns:

Education	Average education of occupational incumbents, years, in 1971.
Income	Average income of incumbents, dollars, in 1971.
Women	Percentage of incumbents who are women.
Prestige	Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.
Census	Canadian census occupational code.
Type	Type of occupation. A factor with levels (note: out of order): bc, Blue Collar; prof, Professional, Managerial, and Technical; wc, White Collar.

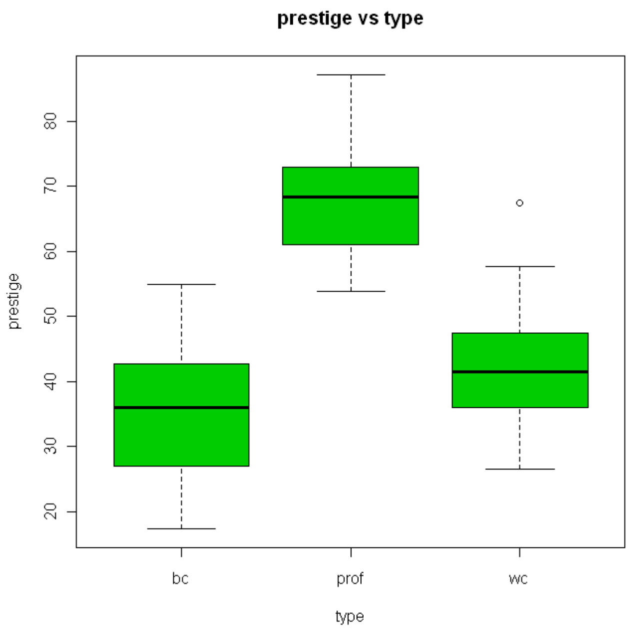
Source

Canada (1971) Census of Canada. Vol. 3, Part 6. Statistics Canada [pp. 19-1–19-21].

3.2-3. ONE-WAY ANOVA: EXAMPLE

First of all, a summary of the data: there are 4 missing values in factor - type

```
> summary(Prestige)
   education          income          women           prestige         census         type
 Min.   : 6.380   Min.   :  611   Min.   : 0.000   Min.   :14.80   Min.   :1113   bc  :44
 1st Qu.: 8.445   1st Qu.: 4106   1st Qu.: 3.592   1st Qu.:35.23   1st Qu.:3120   prof:31
 Median :10.540   Median : 5930   Median :13.600   Median :43.60   Median :5135   wc  :23
                                  Mean   :28.979   Mean   :46.83   Mean   :5402   NA's: 4
                                  3rd Qu.:52.203   3rd Qu.:59.27   3rd Qu.:8312
                                  Max.   :97.510   Max.   :87.20   Max.   :9517
```



- The default R plot to inspect the relation between a numeric variable (prestige) and a factor (type) works well:

```
> plot(prestige~type, main="prestige vs type",col=3)
```

3.2-3. ONE-WAY ANOVA: EXAMPLE

- Descriptive statistics of groups and standard procedure for one-way ANOVA:

```
> tapply(prestige,type,summary)
$bc
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  17.30   27.10   35.90   35.53   42.60   54.90 
$prof
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  53.80   61.00   68.40   67.85   72.95   87.20 
$wc
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  26.50   35.90   41.50   42.24   47.50   67.50 
> oneway.test(prestige~type)
        One-way analysis of means (not assuming equal variances)

F = 117.2471, num df = 2.000, denom df = 54.979, p-value < 2.2e-16

> oneway.test(prestige~type,var=TRUE)# Corresponds to F-Test
        One-way analysis of means

F = 109.5916, num df = 2, denom df = 95, p-value < 2.2e-16

> kruskal.test(prestige~type)# Non Parametric version for One-way means test
        Kruskal-Wallis rank sum test

Kruskal-Wallis chi-squared = 63.3965, df = 2, p-value = 1.713e-14
```

3.2-3. ONE-WAY ANOVA: EXAMPLE

```
> model<-lm(prestige~type, data=Prestige[!is.na(Prestige$type),],
+ contrasts=list(type="contr.treatment"))
> summary(model)

Call:
lm(formula = prestige ~ type, data = Prestige[!is.na(Prestige$type),
    ], contrasts = list(type = "contr.treatment"))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   35.527      1.432  24.810  < 2e-16 ***
typeprof      32.321      2.227  14.511  < 2e-16 ***
typewc         6.716      2.444   2.748  0.00718 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.499 on 95 degrees of freedom
Multiple R-squared: 0.6976,     Adjusted R-squared: 0.6913
F-statistic: 109.6 on 2 and 95 DF,  p-value: < 2.2e-16
```

$$i=1\equiv bc \qquad \hat{y}_{1j}=\bar{y}_1=\hat{\mu}+\hat{\alpha}_1=35.527+0=35.527$$

$$\underbrace{\hat{y}_{ij}=\hat{\mu}+\hat{\alpha}_i}_{\hat{\alpha}_1=0} \rightarrow i=2\equiv prof \;\; \hat{y}_{2j}=\bar{y}_2=\hat{\mu}+\hat{\alpha}_2=35.527+32.321=67.848$$

$$i=3\equiv wc \qquad \hat{y}_{3j}=\bar{y}_3=\hat{\mu}+\hat{\alpha}_3=35.527+6.716=42.244$$

3.2-3. ONE-WAY ANOVA: EXAMPLE

```
> model<-lm(prestige~type, data=Prestige[!is.na(Prestige$type),],
+ contrasts=list(type="contr.sum"))
> summary(model)
Call:
lm(formula = prestige ~ type, data = Prestige[!is.na(Prestige$type),
    ], contrasts = list(type = "contr.sum"))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.5397     0.9935   48.86   <2e-16 ***
type1       -13.0124     1.2925  -10.07   <2e-16 ***
type2        19.3087     1.3990   13.80   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.499 on 95 degrees of freedom
Multiple R-squared: 0.6976,     Adjusted R-squared: 0.6913
F-statistic: 109.6 on 2 and 95 DF,  p-value: < 2.2e-16
```

$$i = 1 \equiv bc \qquad \hat{y}_{1j} = \bar{y}_1 = \hat{\mu} + \hat{\alpha}_1 = 48.540 - 13.013 = 35.527$$

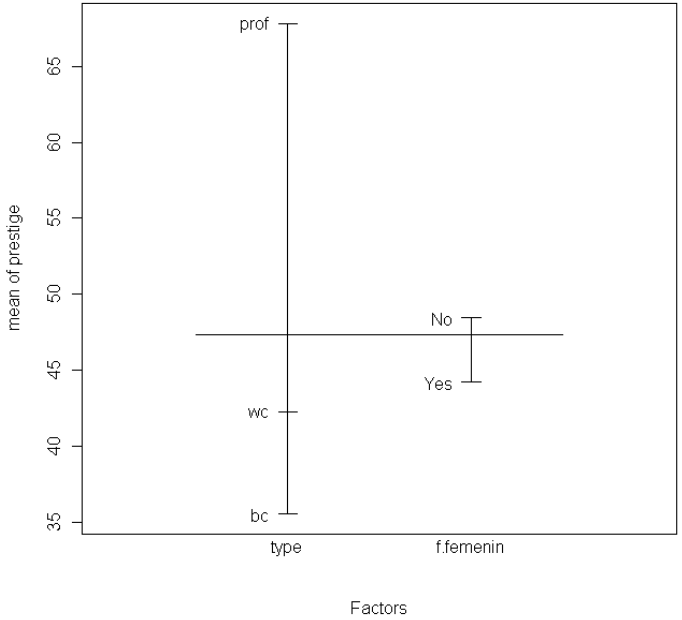
$$\underbrace{\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i}_{\hat{\alpha}_1 + \hat{\alpha}_2 + \hat{\alpha}_3 = 0} \rightarrow i = 2 \equiv prof \qquad \hat{y}_{2j} = \bar{y}_2 = \hat{\mu} + \hat{\alpha}_2 = 48.540 + 19.309 = 67.849$$

$$i = 3 \equiv wc \quad \hat{y}_{3j} = \bar{y}_3 = \hat{\mu} + \hat{\alpha}_3 = 48.540 + (13.013 - 19.309) = 42.244$$

3.2-4 TWO-WAY ANOVA

Motivation: Prestige of professions (Y response) is related to profession type (factor A) and a new factor indicating whether they are mostly female professions (percentage of women greater than 50%) (factor B)?

```
> summary(Prestige)
   education          income           women            prestige          census        type      f.femenin
 Min.   : 6.380   Min.   :  611   Min.   : 0.000   Min.   :14.80   Min.   :1113   bc  :44   No :75
 1st Qu.: 8.445   1st Qu.: 4106   1st Qu.: 3.592   1st Qu.:35.23   1st Qu.:3120   prof:31   Yes:27
 Median :10.540   Median : 5930   Median :13.600   Median :43.60   Median :5135   wc  :23
                                         :28.979   Mean   :46.83   Mean   :5402   NA's: 4
                                      Qu.:52.203   3rd Qu.:59.27   3rd Qu.:8312
                                         :97.510   Max.   :87.20   Max.   :9517
>
```




```
> plot.design(prestige~type+f.femenin)
```

3.2-4. TWO-WAY ANOVA

The analysis of variance of two factors examines the relationship between a quantitative response variable and two explanatory qualitative variables.

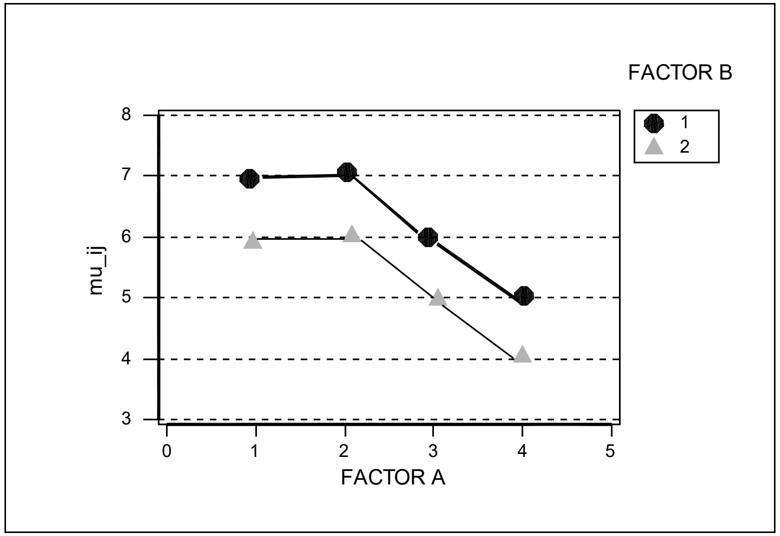
The inclusion of the second factor allows the modeling and standardization of dependence relations and the inclusion of interactions.

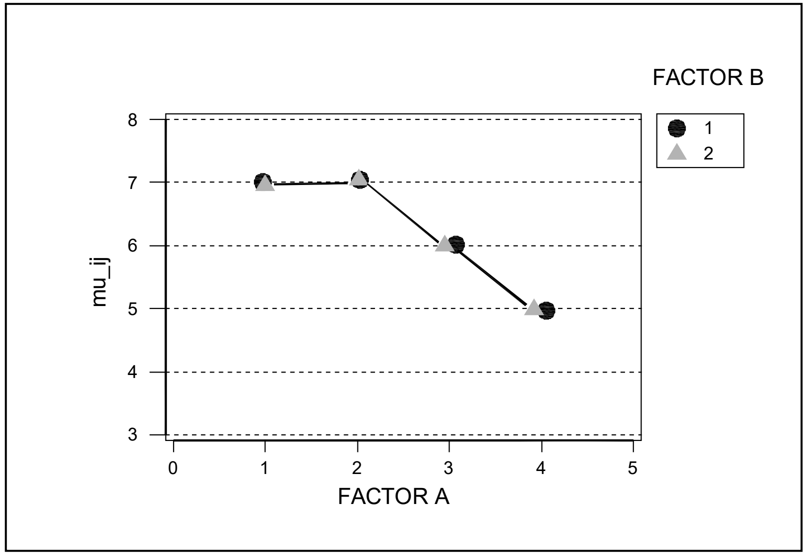
Assuming in a two-way ANOVA that population means for each cell in the combinations of the levels of the factor patterns of usual relationships appear clearly.

A	1	....	J	
1	μ_{11}	....	μ_{1J}	$\mu_{1\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	μ_{I1}	....	μ_{IJ}	$\mu_{I\bullet}$
	$\mu_{\bullet 1}$	....	$\mu_{\bullet J}$	

If factors A and B do not interact, then the partial relationship between each factor and the response variable does not depend on the level of the other factor, that is, the difference between levels is constant. Assume I = 4 and J = 2 in the following diagrams.

3.2-4. TWO-WAY ANOVA





Factors A and B are significant.

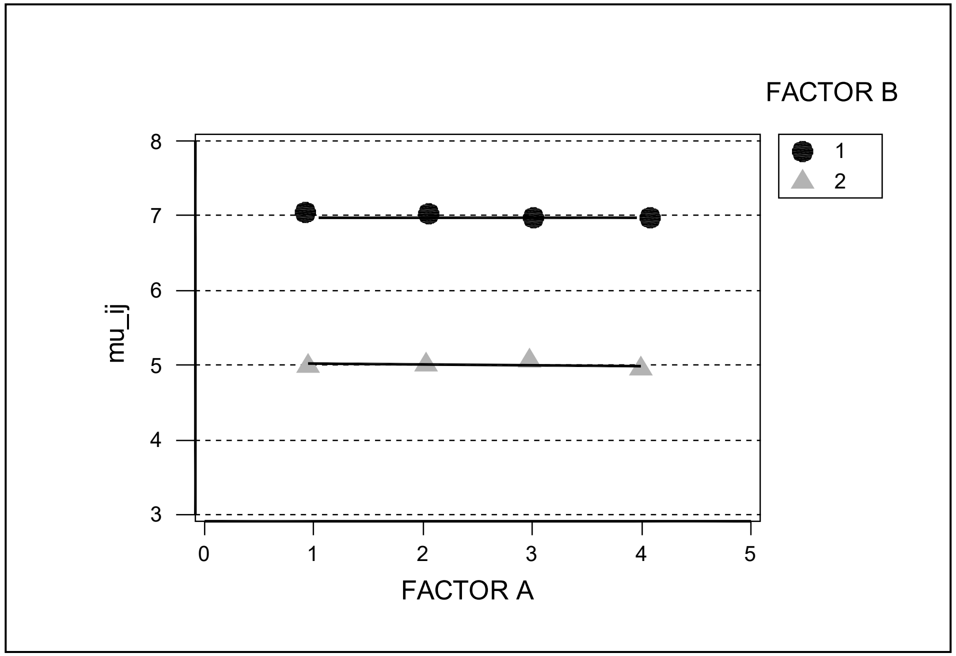
There are no interactions between factors A and B.

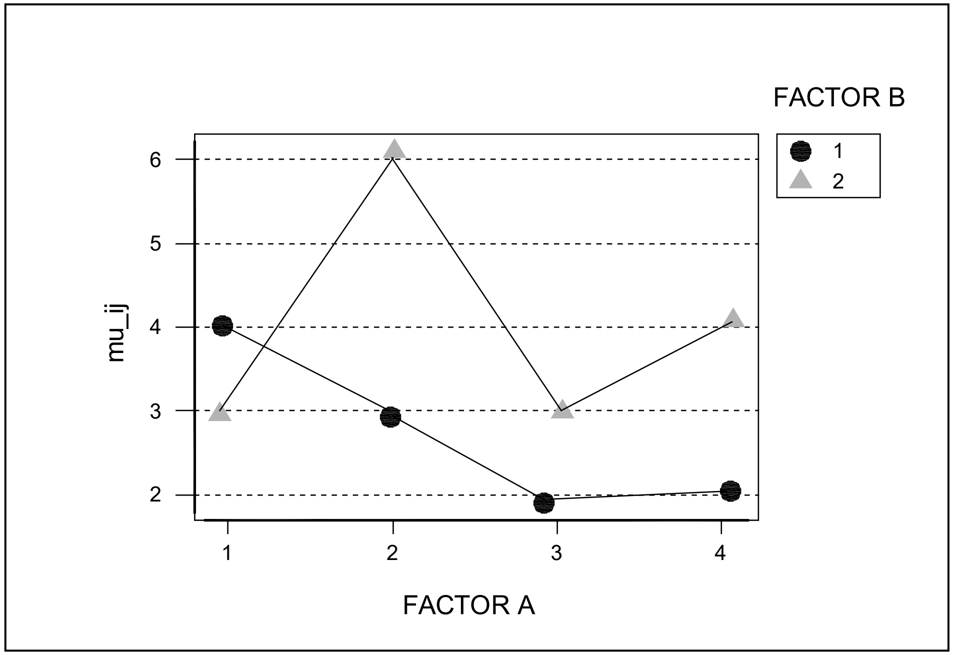
Factor A is significant.

Factor B is not significant.

There are no interactions between factors A and B.

3.2-4. TWO-WAY ANOVA





Factor A is not significant.

Factor B is significant.

There are no interactions between factors A and B.

Factors A and B are significant.

There are interactions between factors A and B.

3.2-4. TWO-WAY ANOVA

Two-way ANOVA models

(M 0)	Null model:	$Y_{ijk} = \mu + \varepsilon_{ijk}$
(M 1)	Two-way ANOVA with interactions:	$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$
(M 2)	Additive two-way ANOVA:	$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$
(M 3)	One-way ANOVA of A:	$Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$
(M 4)	One-way ANOVA of B:	$Y_{ijk} = \mu + \beta_j + \varepsilon_{ijk}$

Most common hypothesis tested in two-way ANOVA: only RSS for models needed.

- **H_1**: There are no interactions between the two factors.

  F-test : H: Model (M1) and (M2) are equivalent `>anova(m2,m1)`

- **H_2**: Net effect of factor A once factor B is included in the model (or factor B | A) is not significant.

  F-test : H: Model (M4) and (M2) are equivalent `>anova(m4,m2)`
  F-test : H: Model (M3) and (M2) are equivalent `>anova(m3,m2)` (for factor B | A)

3.2-4. TWO-WAY ANOVA

- **H_3**: Gross effect of factor A (or factor B).

F-test : H: Model (M0) and (M3) are equivalent `>anova(m0,m3)`
F-test : H: Model (M0) and (M4) are equivalent `>anova(m0,m4)`

3.2-4. TWO-WAY ANOVA

➔ We can give simple mechanical rules to build the design matrix in the model. Assume that **data is ordered according to (levels I, levels J)** for I=J=3 …

➔ The additive model $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$ B factors A and B has a total of 1 + I + J parameters, but only I+J-1 are linearly independent.

➔ Let us build $\mathbf{Y} = \mathbf{X}_R \boldsymbol{\beta}_R + \boldsymbol{\varepsilon}$, when the **baseline reparameterization for level 1** is used.

X=

1	1			1		
1	1				1	
1	1					1
1		1		1		
1		1			1	
1		1				1
1			1	1		
1			1		1	
1			1			1
μ	α_1	...	α_I	β_1	...	β_J

XR =

1			0	0
1			1	0
1			0	1
1	1		0	0
1	1		1	0
1	1		0	1
1		1	0	0
1		1	1	0
1		1	0	1
μ	α_2	α_I	β_2	β_J

$$\boxed{\begin{aligned} \alpha_1 &= 0 \\ \beta_1 &= 0 \end{aligned}}$$

3.2-4. TWO-WAY ANOVA

- ➔ The complete two-way ANOVA model with interactions $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$ for factors A and B has a total of 1 + I + J+I*J parameters, but only I*J are linearly independent.
- ➔ Let us build $\mathbf{Y} = \mathbf{X}_R \boldsymbol{\beta}_R + \boldsymbol{\varepsilon}$ for both zero sum and **baseline (ref 1) reparameterization**.

X_R =

μ	α_1	α_{I-1}	β_1	β_{J-1}	γ_{11}	$\gamma_{1,J-1}$	$\gamma_{I-1,1}$	$\gamma_{I-1,J-1}$
			1		1			
1	1			1		1		
			-1	-1	-1	-1		
			1				1	
1		1		1				1
			-1	-1			-1	-1
			1		-1		-1	
1	-1	-1		1		-1		-1
			-1	-1	1	1	1	1

Zero Sum

X_R =

μ	α_2	α_I	β_2	β_J	γ_{22}	γ_{23}	γ_{32}	γ_{33}
1			0	0	0	0	0	0
1			1	0	0	0	0	0
1			0	1	0	0	0	0
1	1		0	0	0	0	0	0
1	1		1	0	1	0	0	0
1	1		0	1	0	1	0	0
1		1	0	0	0	0	0	0
1		1	1	0	0	0	1	0
1		1	0	1	0	0	0	1

Base-line i=j=1

3.2-4. TWO-WAY ANOVA

➔ For baseline $\alpha_1 = 0$ and $\beta_1 = 0$ plus , $\gamma_{1j} = 0 \quad \forall j = 1 \ldots J$, $\gamma_{i1} = 0 \ \forall\, i = 1 \cdots I$, with *I+J* restrictions, one of which is redundant, let us assume the first is $\gamma_{11} = 0$.

$\gamma_{11} = 0$	$\cdots$	$\gamma_{1,J-1} = 0$	$\gamma_{1J} = 0$
$\gamma_{21} = 0$	$\cdots$	$\gamma_{2,J-1}$	γ_{2J}
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\gamma_{I1} = 0$	$\cdots$	$\gamma_{I,J-1}$	γ_{IJ}

3.2-4. TWO-WAY ANOVA: EXAMPLE

Example: Prestige of Canadian occupations vs type and female factor (Fox and Weisberg 2011)

```
> options(contrasts=c("contr.treatment","contr.treatment"))
> m0<-lm(prestige~1,data=Prestige[!is.na(Prestige$type),])
> m1<-lm(prestige~type*f.femenin,data=Prestige[!is.na(Prestige$type),])
> m2<-lm(prestige~type+f.femenin,data=Prestige[!is.na(Prestige$type),])
> m3<-lm(prestige~type,data=Prestige[!is.na(Prestige$type),])
> m4<-lm(prestige~f.femenin,data=Prestige[!is.na(Prestige$type),])
> anova(m2,m1) # Interaction Test
Analysis of Variance Table
Model 1: prestige ~ type + f.femenin
Model 2: prestige ~ type * f.femenin
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     94 8382.2
2     92 8194.6  2    187.61 1.0531  0.353

> anova(m3,m2) # Net f.femenin effect
Analysis of Variance Table
Model 1: prestige ~ type
Model 2: prestige ~ type + f.femenin
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     95 8571.3
2     94 8382.2  1    189.07 2.1203 0.1487
```

3.2-4. TWO-WAY ANOVA

```
> anova(m4,m2) # Net type effect

Analysis of Variance Table
Model 1: prestige ~ f.femenin
Model 2: prestige ~ type + f.femenin
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     96 28001.6
2     94  8382.2  2     19619 110.01 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(m0,m3) # Gross type effect

Analysis of Variance Table
Model 1: prestige ~ 1
Model 2: prestige ~ type
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     97 28346.9
2     95  8571.3  2     19776 109.59 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.2-4. TWO-WAY ANOVA

```
> anova(m0,m4) # Gross f.femenin effect
Analysis of Variance Table

Model 1: prestige ~ 1
Model 2: prestige ~ f.femenin
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     97 28347
2     96 28002  1    345.31 1.1838 0.2793
> summary(m1)
         Y(type="wc", f.femenin="Yes")= 36.23 + 5.46 – 4.4 + 5.38 = 42.678
Call:
lm(formula = prestige ~ type*f.femenin, data=Prestige[!is.na(Prestige$type),    ])

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             36.227      1.552  23.349   <2e-16 ***
typeprof                33.033      2.443  13.519   <2e-16 ***
typewc                   5.463      3.364   1.624    0.108
f.femeninYes            -4.398      3.890  -1.131    0.261
typeprof:f.femeninYes   -2.895      5.791  -0.500    0.618
typewc:f.femeninYes      5.378      5.558   0.968    0.336

Residual standard error: 9.438 on 92 degrees of freedom
Multiple R-squared: 0.7109,     Adjusted R-squared: 0.6952
F-statistic: 45.25 on 5 and 92 DF,  p-value: < 2.2e-16
```

3.2-4. TWO-WAY ANOVA

```
> summary(m2)
Call:
lm(formula = prestige ~ type + f.femenin,data=Prestige[!is.na(Prestige$type),])

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    36.068      1.471  24.516  < 2e-16 ***
typeprof       32.438      2.216  14.640  < 2e-16 ***
typewc          8.096      2.608   3.104  0.00252 **
f.femeninYes   -3.398      2.333  -1.456  0.14869

Residual standard error: 9.443 on 94 degrees of freedom
Multiple R-squared: 0.7043,     Adjusted R-squared: 0.6949
F-statistic: 74.63 on 3 and 94 DF,  p-value: < 2.2e-16

>
```

3.2-4. TWO-WAY ANOVA

```
> options(contrasts=c("contr.sum","contr.sum"))
> m2<-lm(prestige~type+f.femenin,data=Prestige[!is.na(Prestige$type),])
> summary(m2)
Call:lm(formula = prestige ~ type + f.femenin,data=Prestige[!is.na(Prestige$type),])
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   47.880      1.087  44.066   <2e-16 ***
type1        -13.511      1.330 -10.160   <2e-16 ***
type2         18.927      1.415  13.372   <2e-16 ***
f.femenin1     1.699      1.167   1.456    0.149
---
Residual standard error: 9.443 on 94 degrees of freedom
Multiple R-squared: 0.7043,     Adjusted R-squared: 0.6949
F-statistic: 74.63 on 3 and 94 DF,  p-value: < 2.2e-16>
```

3.2-4. TWO-WAY ANOVA

```
> options(contrasts=c("contr.sum","contr.sum"))
> m1<-lm(prestige~type*f.femenin,data=Prestige[!is.na(Prestige$type),])
> summary(m1)
Call:lm(formula = prestige ~ type * f.femenin, data =
Prestige[!is.na(Prestige$type),])
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       47.2736     1.1702  40.397  < 2e-16 ***
type1            -13.2458     1.6219  -8.167 1.62e-12 ***
type2             18.3398     1.7039  10.763  < 2e-16 ***
f.femenin1         1.7854     1.1702   1.526    0.131
type1:f.femenin1   0.4138     1.6219   0.255    0.799
type2:f.femenin1   1.8612     1.7039   1.092    0.278
---
Residual standard error: 9.438 on 92 degrees of freedom
Multiple R-squared:  0.7109,  Adjusted R-squared:  0.6952
F-statistic: 45.25 on 5 and 92 DF,  p-value: < 2.2e-16
```

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \text{ with } \alpha_1 + \alpha_2 + \alpha_3 = 0, \beta_1 + \beta_2 = 0, \gamma_{1j} + \gamma_{2j} + \gamma_{3j} = 0 \text{ and } \gamma_{i1} + \gamma_{i2} = 0$$

i=1 and j=1: $Y_{11} = \mu + \alpha_1 + \beta_1 + \gamma_{11} =$ 47.27 -13.25 +1.79 +0.41 = 36.22
i=2 and j=1: $Y_{21} = \mu + \alpha_2 + \beta_1 + \gamma_{21} =$ 47.27 +18.34 +1.79 +1.86
i=3 and j=1: $Y_{31} = \mu + \alpha_3 + \beta_1 + \gamma_{31} =$ 47.27 +13.25-18.24 +1.79 -0.41-1.86
i=1 and j=2: $Y_{12} = \mu + \alpha_1 + \beta_2 + \gamma_{12} =$ 47.27 -13.25 -1.79 -0.41
i=2 and j=2: $Y_{22} = \mu + \alpha_2 + \beta_2 + \gamma_{22} =$ 47.27 +18.34 -1.79 -1.86
i=3 and j=2: $Y_{32} = \mu + \alpha_3 + \beta_2 + \gamma_{32} =$ 47.27 +13.25-18.24 -1.79 +0.41+1.86 = 42.7

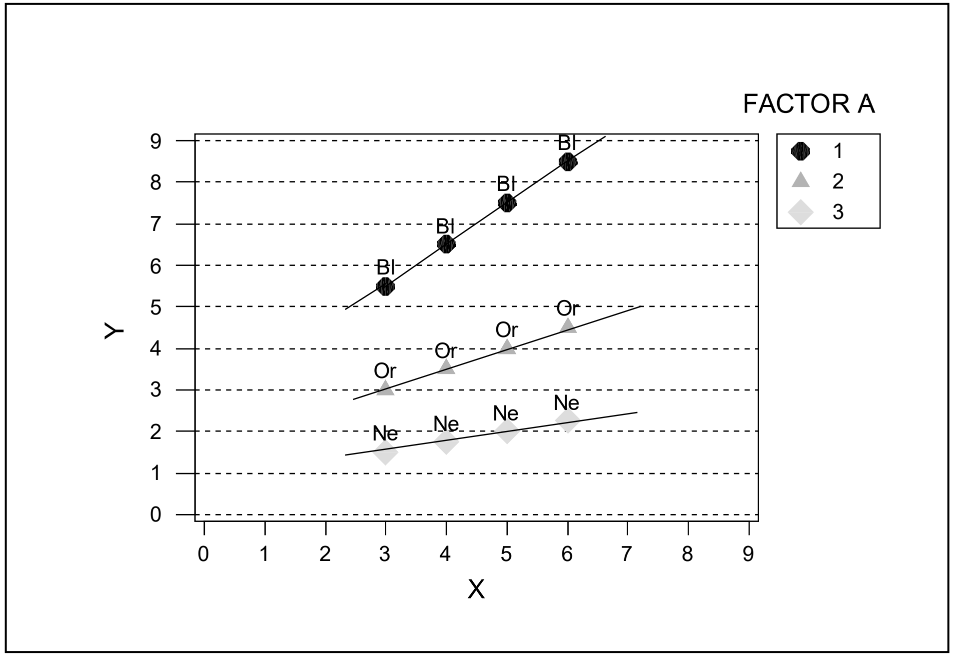
3.2-5 MORE COMPLEX ANOVA MODELS

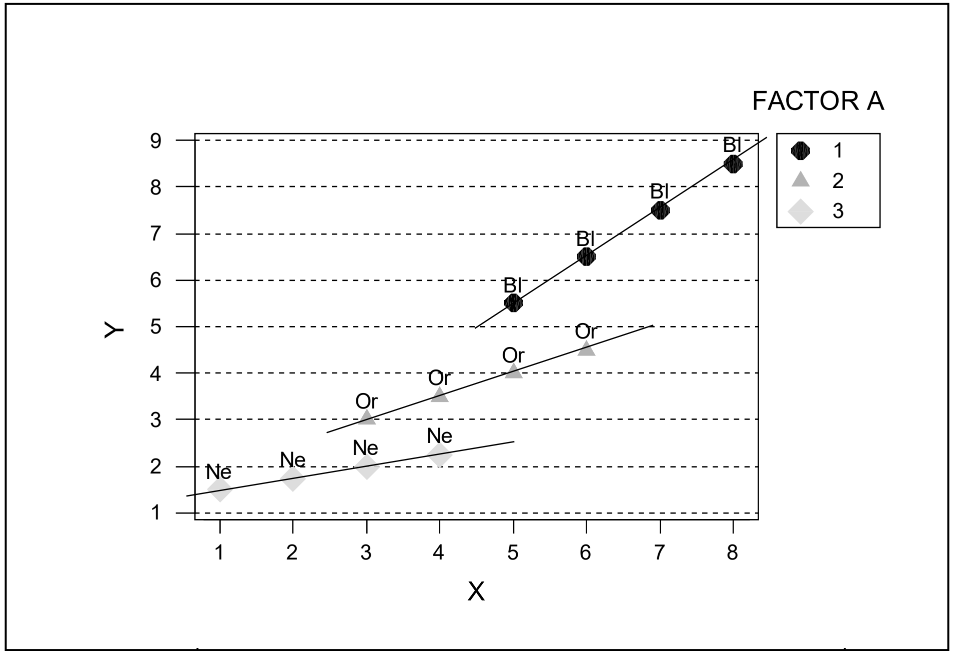
Extending the formula to more complex ANOVA models is straightforward, for example, when increasing the number of factors in the experimental designs. When the number of factors is increased to factors A, B and C, then higher-order interactions appear, the three-order interaction A*B*C and 3 two-order interactions A*B, A*C, B*C and the main effects. Interpretation of model coefficients gets complex.

In the designs of real experiments, the factors can be crossed or nested or both: they can all be easily incorporated.

3.2-6 ANALYSIS OF COVARIANCE (ANCOVA MODELS)

- ANCOVA models of analysis of covariance are mixed models that have dummy variables that represent levels of factors or interactions and continuous variables or covariates.
- We are going to see ANCOVA models through an example without data, dealing with sociology and which is very intuitive, inspired by the Fox (1984) proposal: relation between income (Y) and level of education (X) among the white, oriental and black populations of the US (factor A, I=3).



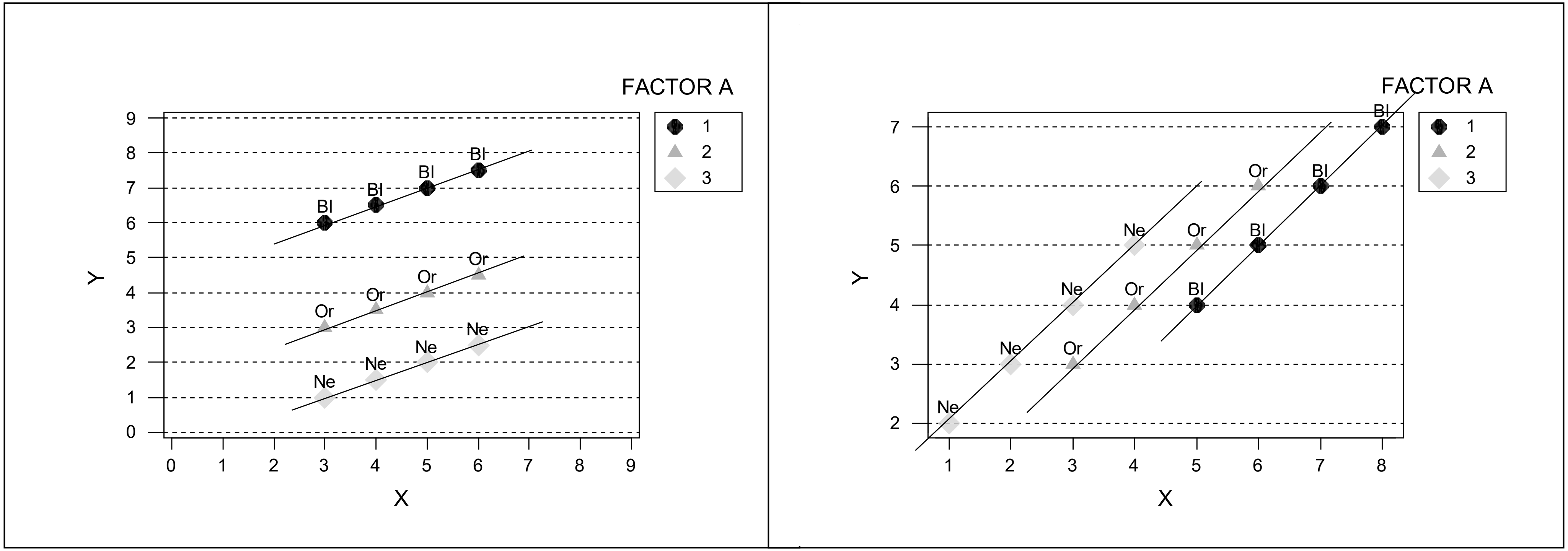


$$Y_{ik} = \mu + \alpha_i + (\eta + \theta_i)\, x_{ik} + \varepsilon_{ik}$$

Model (M1):

Factor-covariate interaction

3.2-6. ANALYSIS OF COVARIANCE (ANCOVA MODELS)



Model (M2):

No factor-covariate interaction, no correlation between race and education.

$$Y_{ik} = \mu + \alpha_i + \eta\, x_{ik} + \varepsilon_{ik}$$

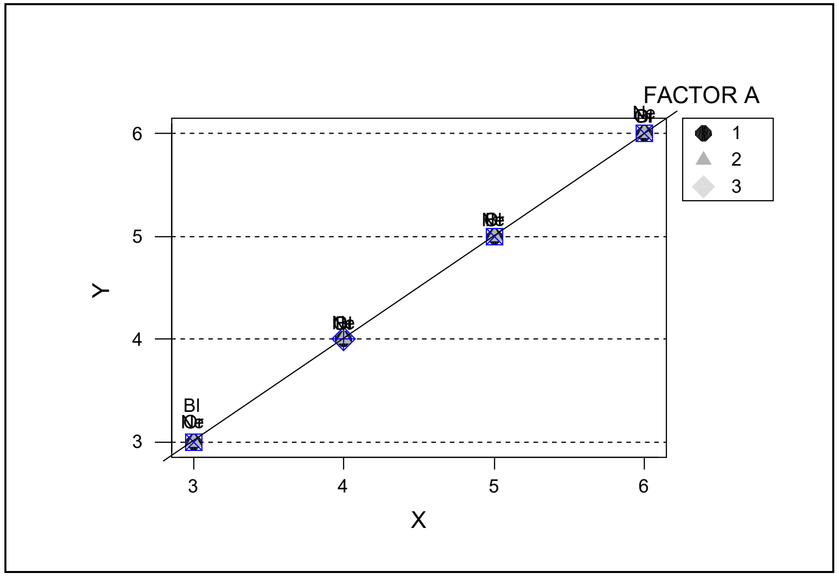
Model (M2):

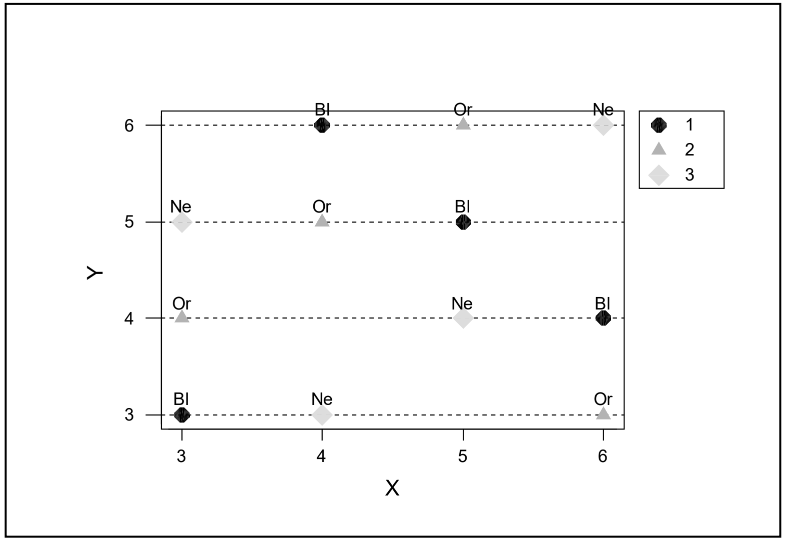
No factor-covariate interaction, correlation between race and education.

3.2-6. ANALYSIS OF COVARIANCE (ANCOVA MODELS)

Model (M4) Income and education associated with but not affected by race

$$Y_{ik} = \mu + \eta\, x_{ik} + \varepsilon_{ik} \Longrightarrow$$





Model (M4) Null model

$$Y_{ik} = \mu + \varepsilon_{ik}$$

3.2-6. ANALYSIS OF COVARIANCE (ANCOVA MODELS)

(M 2)

Additive ANCOVA model (same slope) $Y_{ik} = \mu + \alpha_i + \eta\, x_{ik} + \varepsilon_{ik}$, has 5 *(=1+3+1)* *parameters* *under baseline i=1*, independent columns *after forcing*.

$$\alpha_1 = 0$$

X=

1	1			x1
1		1		x2
1			1	x3
.1.	α_1	α_2	α_3	η

XR =

1	0	0	x1
1	1	0	x2
1	0 0 0	1 1 1	x3
μ	α_2	α_3	η

X= XR =

1	x1
1	x2
1	x3
μ	η

The simple regression model $Y_{ik} = \alpha + \eta\, x_{ik} + \varepsilon_{ik}$ has *2 (=1+1) parameters.*

(M 3)

3.2-6. ANALYSIS OF COVARIANCE (ANCOVA MODELS)

(M 1)

The ANCOVA model with interactions (different slope)

$Y_{ik} = \mu + \alpha_i + (\eta + \theta_i)\, x_{ik} + \varepsilon_{ik}$, has 8 (=1+3+4) parameters $\boxed{\text{under baseline } i=1}$ *reparameterization leads to 6 (=1+2+1+2)* independent columns *after forcing.*

X=

1	1			x1	x1		
1		1		x2		x2	
1			1	x3			x3
.1.	α_1	α_2	α_3	η	θ_1	θ_2	.1.:

X_R =

1	0 0 0	0 0 0	x1	0	0
1	1 1 1	0 0 0	x2	x2	0
1	0 0 0	1 1 1	x3	0 0 0	0 x3 0
μ	α_2	α_3	η	θ_2	θ_3

$$\begin{array}{|lcr|} \alpha_1 & = & 0 \\ \theta_1 & = & 0 \end{array}$$

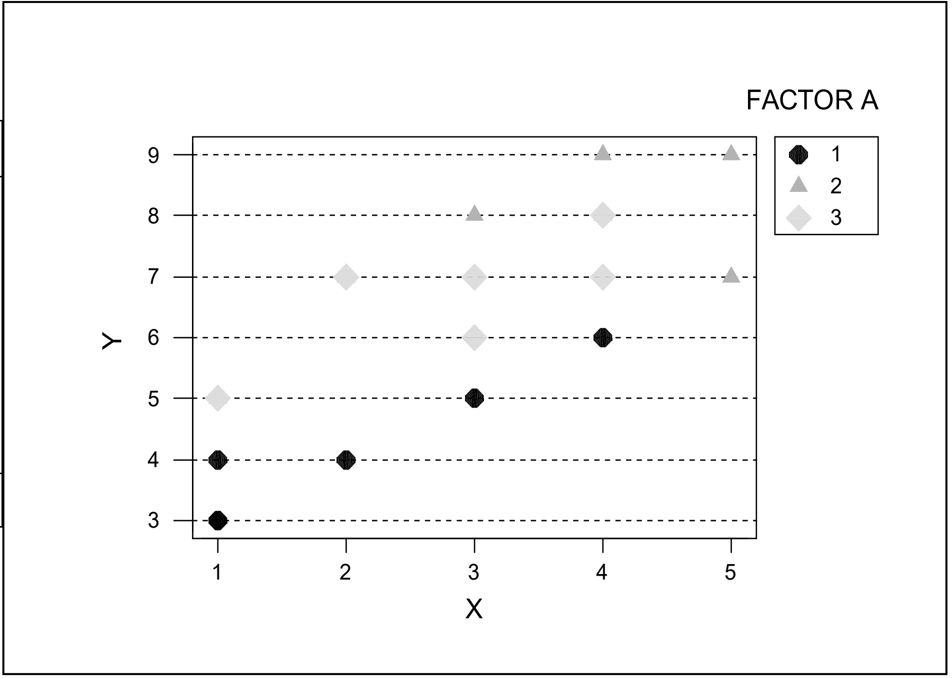
ANALYSIS OF COVARIANCE (ANCOVA MODELS)

Manual example: Training type for runners (Dobson 1990)

The data show the results obtained by 21 runners as three levels of a factor, representing three different methods of training, and an explanatory variable (covariate) that represents the score obtained before initiating the training. We want to compare the methods of training taking into account the different initial capacities (Dobson, 1990).

Factor A

Reply	A_1		A_2		A_3	
k=1	6	3	8	4	6	3
k=2	4	1	9	5	7	2
k=3	5	3	7	5	7	2
k=4	3	1	9	4	7	3
k=5	4	2	8	3	8	4
k=6	3	1	5	1	5	1
k=7	6	4	7	2	7	4
(y, x)	y	x	y	x	y	x



3.2-7. ANALYSIS OF COVARIANCE: MANUAL EXAMPLE

$$\begin{array}{c} 11\\12\\13\\14\\15\\16\\17\\21\\22\\23\\24\\25\\26\\27\\31\\32\\33\\34\\35\\36\\37 \end{array}
\underbrace{\begin{pmatrix}6\\4\\5\\3\\4\\3\\6\\8\\9\\7\\9\\8\\5\\7\\6\\7\\7\\7\\8\\5\\7\end{pmatrix}}_{\mathbf{Y}}
=
\underbrace{\begin{pmatrix}
1&0&0&3&0&0\\
1&0&0&1&0&0\\
1&0&0&3&0&0\\
1&0&0&1&0&0\\
1&0&0&2&0&0\\
1&0&0&1&0&0\\
1&0&0&4&0&0\\
1&1&0&4&4&0\\
1&1&0&5&5&0\\
1&1&0&5&5&0\\
1&1&0&4&4&0\\
1&1&0&3&3&0\\
1&1&0&1&1&0\\
1&1&0&2&2&0\\
1&0&1&3&0&3\\
1&0&1&2&0&2\\
1&0&1&2&0&2\\
1&0&1&3&0&3\\
1&0&1&4&0&4\\
1&0&1&1&0&1\\
1&0&1&4&0&4
\end{pmatrix}}_{\mathbf{X_R}}
\underbrace{\begin{pmatrix}\mu\\ \alpha_2\\ \alpha_3\\ \eta\\ \theta_2\\ \theta_3\end{pmatrix}}_{\boldsymbol{\beta}_R}
+
\underbrace{\begin{pmatrix}\varepsilon_{11}\\ \varepsilon_{12}\\ \varepsilon_{13}\\ \varepsilon_{14}\\ \varepsilon_{15}\\ \varepsilon_{16}\\ \varepsilon_{17}\\ \varepsilon_{21}\\ \varepsilon_{22}\\ \varepsilon_{23}\\ \varepsilon_{24}\\ \varepsilon_{25}\\ \varepsilon_{26}\\ \varepsilon_{27}\\ \varepsilon_{31}\\ \varepsilon_{32}\\ \varepsilon_{33}\\ \varepsilon_{34}\\ \varepsilon_{35}\\ \varepsilon_{36}\\ \varepsilon_{37}\end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

➔ The complete ANCOVA has 8 (=1+3+4) parameters $Y_{ik}=\mu+\alpha_i+(\eta+\theta_i)x_{ik}+\varepsilon_{ik}$ plus the baseline i=1 restrictions:

$$\boxed{\begin{array}{lcl}\alpha_1 &=& 0\\ \theta_1 &=& 0\end{array}}$$

$$\mathbf{b_R} = \left(\mathbf{X_R^T X_R}\right)^{-1}\mathbf{X_R}^{\mathbf{T}}\mathbf{Y} = \begin{pmatrix}2{,}35\\ 2{,}9\\ 2{,}74\\ 0{,}968\\ -0{,}151\\ -0{,}368\end{pmatrix}$$

$$\hat{y}_{1.} = \hat{\mu}+\hat{\alpha}_1+\left(\hat{\eta}+\hat{\theta}_1\right)x = \hat{\mu}+0+\left(\hat{\eta}+0\right)x = 2{,}35+0{,}968x$$
$$\hat{y}_{2.} = \hat{\mu}+\hat{\alpha}_2+\left(\hat{\eta}+\hat{\theta}_2\right)x = 2{,}35+2{,}9+\left(0{,}968-0{,}151\right)x = 5{,}25+0{,}817x$$
$$\hat{y}_{3.} = \hat{\mu}+\hat{\alpha}_3+\left(\hat{\eta}+\hat{\theta}_3\right)x = 2{,}35+2{,}74+\left(0{,}968-0{,}368\right)x = 5{,}09+0{,}6x$$

3.2-7. ANALYSIS OF COVARIANCE: MANUAL EXAMPLE

$$
\begin{array}{c}
11\\12\\13\\14\\15\\16\\17\\21\\22\\23\\24\\25\\26\\27\\31\\32\\33\\34\\35\\36\\37
\end{array}
\underbrace{\begin{pmatrix}6\\4\\5\\3\\4\\3\\6\\8\\9\\7\\9\\8\\5\\7\\6\\7\\7\\7\\8\\5\\7\end{pmatrix}}_{\mathbf{Y}}
=
\underbrace{\begin{pmatrix}
1&0&0&3\\
1&0&0&1\\
1&0&0&3\\
1&0&0&1\\
1&0&0&2\\
1&0&0&1\\
1&0&0&4\\
1&1&0&4\\
1&1&0&5\\
1&1&0&5\\
1&1&0&4\\
1&1&0&3\\
1&1&0&1\\
1&1&0&2\\
1&0&1&3\\
1&0&1&2\\
1&0&1&2\\
1&0&1&3\\
1&0&1&4\\
1&0&1&1\\
1&0&1&4
\end{pmatrix}}_{\mathbf{X_R}}
\underbrace{\begin{pmatrix}\mu\\ \alpha_2\\ \alpha_3\\ \eta\end{pmatrix}}_{\boldsymbol{\beta}_{\mathbf{R}}}
+\boldsymbol{\varepsilon}
$$

Additive ANCOVA $Y_{ik} = \mu + \alpha_i + \eta x_{ik} + \varepsilon_{ik}$ plus the **baseline i=1 restriction**

$$\boxed{\alpha_1 = 0}$$

$$
\boxed{
\begin{array}{l}
\hat{\mu} = b_1 = 2{,}846 \\
\hat{\alpha}_2 = b_2 = 2{,}188 \quad \hat{\alpha}_3 = b_3 = 1{,}862 \quad \hat{\alpha}_1 = 0 \\
\hat{\eta} = b_4 = 0{,}743
\end{array}
}
$$

$$
\mathbf{X_R^T X_R\, b_R} = \mathbf{X_R}^{\mathbf{T}}\mathbf{Y} \Leftrightarrow \mathbf{b_R} = \left(\mathbf{X_R^T X_R}\right)^{-1}\mathbf{X_R^T Y} = \begin{pmatrix}2{,}846\\2{,}188\\1{,}862\\0{,}743\end{pmatrix}
$$

3.2-7. ANALYSIS OF COVARIANCE: MANUAL EXAMPLE

Example: Prestige of Canadian occupations in data.frame Prestige in car library for R (Fox and Weisberg 2011)

An ANCOVA model in which income and type of profession is tested for the prediction of prestige illustrates how to estimate and contrast hypotheses (using the F test) in R. Interpretation has to be considered

```
> summary(Prestige)
   education          income          women            prestige         census         type     f.femenin
 Min.   : 6.380   Min.   :  611   Min.   : 0.000   Min.   :14.80   Min.   :1113   bc  :44   No :75
 1st Qu.: 8.445   1st Qu.: 4106   1st Qu.: 3.592   1st Qu.:35.23   1st Qu.:3120   prof:31   Yes:27
 Median :10.540   Median : 5930   Median :13.600   Median :43.60   Median :5135   wc  :23
 Mean   :10.738   Mean   : 6798   Mean   :28.979   Mean   :46.83   Mean   :5402   NA's: 4
 3rd Qu.:12.648   3rd Qu.: 8187   3rd Qu.:52.203   3rd Qu.:59.27   3rd Qu.:8312
 Max.   :15.970   Max.   :25879   Max.   :97.510   Max.   :87.20   Max.   :9517

> options(contrasts=c("contr.treatment","contr.treatment"))
> m0<-lm(prestige~1,data=Prestige[!is.na(Prestige$type),])
> m1<-lm(prestige~type*income,data=Prestige[!is.na(Prestige$type),])
> m2<-lm(prestige~type+income,data=Prestige[!is.na(Prestige$type),])
> m3<-lm(prestige~type,data=Prestige[!is.na(Prestige$type),])
> m4<-lm(prestige~income,data=Prestige[!is.na(Prestige$type),])
>
```

3.2-7. ANALYSIS OF COVARIANCE: PRESTIGE EXAMPLE

```
> anova(m2,m1) # Interaction Test
Analysis of Variance Table
Model 1: prestige ~ type + income
Model 2: prestige ~ type * income
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     94 6336.7
2     92 4859.2  2    1477.5 13.987 4.969e-06 ***
---
> anova(m3,m2) # Net income-covariate effect
Analysis of Variance Table
Model 1: prestige ~ type
Model 2: prestige ~ type + income
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     95 8571.3
2     94 6336.7  1    2234.5 33.147 1.068e-07 ***
---

> anova(m4,m2) # Net type effect
Analysis of Variance Table
Model 1: prestige ~ income
Model 2: prestige ~ type + income
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     96 14325.3
2     94  6336.7  2    7988.5 59.251 < 2.2e-16 ***
---
```

3.2-7. ANALYSIS OF COVARIANCE: PRESTIGE EXAMPLE

```
> anova(m0,m3) # Gross income-covariate effect
Analysis of Variance Table
Model 1: prestige ~ 1
Model 2: prestige ~ type
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     97 28346.9
2     95  8571.3  2     19776 109.59 < 2.2e-16 ***
---

> anova(m0,m4) # Gross type effect
Analysis of Variance Table

Model 1: prestige ~ 1
Model 2: prestige ~ income
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     97 28347
2     96 14325  1     14022 93.965 6.773e-16 ***
---

> summary(m1)

Call:lm(formula =prestige~type*income, data=Prestige[!is.na(Prestige$type),])
```

3.2-7. ANALYSIS OF COVARIANCE: PRESTIGE EXAMPLE

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     13.9045168  3.1671787   4.390 3.02e-05 ***
typeprof        45.0190221  4.2907398  10.492  < 2e-16 ***
typewc          18.9807386  5.3421020   3.553 0.000603 ***
income           0.0040235  0.0005530   7.276 1.12e-10 ***
typeprof:income -0.0031783  0.0006047  -5.256 9.48e-07 ***
typewc:income   -0.0021712  0.0009700  -2.238 0.027603 *  
---

Residual standard error: 7.268 on 92 degrees of freedom
Multiple R-squared: 0.8286,     Adjusted R-squared: 0.8193 
F-statistic: 88.94 on 5 and 92 DF,  p-value: < 2.2e-16
```

$$Y_{ik} = \mu + \alpha_i + (\eta + \theta_i)x_{ik} \ \ with\ \alpha_1 = 0\ and\ \theta_1 = 0$$

i=1: $Y_{1k} = \mu + \alpha_1 + (\eta + \theta_1)x_{1k} = (13.91 + 0) + (0.0040 + 0)x_{1k}$
i=2: $Y_{2k} = \mu + \alpha_2 + (\eta + \theta_2)x_{2k} = (13.91 + 45.02) + (0.0040 - 0.0032)x_{2k}$
i=3: $Y_{3k} = \mu + \alpha_3 + (\eta + \theta_3)x_{3k} = (13.91 + 18.98) + (0.0040 - 0.0022)x_{3k}$

3.2-7. ANALYSIS OF COVARIANCE: PRESTIGE EXAMPLE

```
> options(contrasts=c("contr.sum","contr.sum"))
> m1<-lm(prestige~type*income,data=Prestige[!is.na(Prestige$type),])
> summary(m1)
Call:lm(formula =prestige~type*income, data=Prestige[!is.na(Prestige$type),])
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.524e+01  2.025e+00  17.399  < 2e-16 ***
type1        -2.133e+01  2.729e+00  -7.818 8.59e-12 ***
type2         2.369e+01  2.626e+00   9.020 2.63e-14 ***
income        2.240e-03  3.334e-04   6.719 1.50e-09 ***
type1:income  1.783e-03  4.616e-04   3.863 0.000208 ***
type2:income -1.395e-03  3.621e-04  -3.852 0.000216 ***
---
Residual standard error: 7.268 on 92 degrees of freedom
Multiple R-squared: 0.8286,     Adjusted R-squared: 0.8193
F-statistic: 88.94 on 5 and 92 DF,  p-value: < 2.2e-16
```

$$Y_{ik} = \mu + \alpha_i + (\eta + \theta_i)x_{ik} \ \ with\ \alpha_1 + \alpha_2 + \alpha_3 = 0\ and\ \theta_1 + \theta_2 + \theta_3 = 0$$

i=1: $Y_{1k} = \mu + \alpha_1 + (\eta + \theta_1)x_{1k} = (35.24 - 21.33) + (0.0022 + 0.0018)x_{1k}$

i=2: $Y_{2k} = \mu + \alpha_2 + (\eta + \theta_2)x_{2k} = (35.24 + 23.69) + (0.0022 - 0.0014)x_{2k}$

i=3: $Y_{3k} = \mu + \alpha_3 + (\eta + \theta_3)x_{3k} = (35.24 + 21.33 - 23.69) + (0.0022 - 0.0018 + 0.0014)x_{3k}$